Josep Francesc Abril Ferrando

Comparative Analysis of Eukaryotic Gene Sequence Features

Anàlisi Comparativa d'Elements de Seqüència dels Gens Eucariotes

PhD Thesis

Barcelona, May 2005

Dipòsit legal: B.47269-2005
ISBN: 978-84-691-1209-0

Comparative Analysis of Eukaryotic Gene Sequence Features

Anàlisi Comparativa d'Elements de Seqüència dels Gens Eucariotes

Josep Francesc Abril Ferrando

PhD Thesis

Barcelona, May 2005

First Edition, April 2005.

Printed at:

COPISTERIA MIRACLE
Rector Ubach, 6–10 (Aribau corner)
08021 — Barcelona
Phone: +034 93 200 85 44
Fax: +034 93 209 17 82
Email: `miracle at miraclepro.com`

Cover Figure:

An artistic representation of how Bioinformatics helped to decode the human genome. Metaphasic chromosomes are lying on top of a changing background where the DNA nucleic acids—A, C, G, and T, the language of life—, are converted into a binary code—0's and 1's, the language of computers—. A montage by J.F. Abril made with the `Gimp` (`http://www.gimp.org/`).

UNIVERSITAT POMPEU FABRA
Departament de Ciències Experimentals i de la Salut

Comparative Analysis of Eukaryotic Gene Sequence Features

Anàlisi Comparativa d'Elements de Seqüència dels Gens Eucariotes

Josep Francesc Abril Ferrando

Memòria presentada per optar al grau de Doctor
en Biologia per la Universitat Pompeu Fabra.

Aquesta Tesi Doctoral ha estat realitzada sota la direcció del
Dr. *Roderic Guigó i Serra* al Departament de Ciències Experimentals
i de la Salut de la Universitat Pompeu Fabra.

Roderic Guigó i Serra **Josep Francesc Abril Ferrando**

Barcelona, May 2005

The research in this thesis has been carried out at the Genome BioInformatics Lab (GBIL) within the Grup de Recerca en Informàtica Biomèdica (GRIB) at the Parc de Recerca Biomèdica de Barcelona (PRBB), a consortium of the Institut Municipal d'Investigació Mèdica (IMIM), the Universitat Pompeu Fabra (UPF) and the Centre de Regulació Genòmica (CRG).

The research carried out in this thesis has been supported by predoctoral fellowships from Instituto de Salut Carlos III (Beca de Formación de personal Investigador, BEFI, 1999-2003) and from Fundació IMIM (2003-2004) to J. F. Abril, and grants from Ministerio de Ciencia y Tecnología to R. Guigó.

To my wife Marta,
    for her ever lasting patience
    with me and computers...

To my daughter Ruth,
    for taking all those dark clouds
    away with her smiles...

Preface

During the last century biologists have been accumulating an overwhelming amount of information, but it has been during the last decade when we have experienced an explosion of data acquisition. At all levels, living beings have become more and more complex than the reductionists would ever have expected. Never before it was possible to assert, as nowadays, that life is not only the sum of the constituent molecules, acting as the gears of a clock, but also the raising network of interactions between them. Biology, starting as a descriptive subject, has evolved into an information-driven subject, taking biologists from the wet lab to the computer screens. Currently, quoting Lincoln Stein from his foreword to Tisdall [2003], "if you can't do Bioinformatics, you can't do Biology".

We, as humans, are prone to define sets, clustering elements with similar features into groups, to face the complexity. Within this landscape, a bunch of "omics" terms have been coined. We will focus on the analysis of genomic sequences, more precisely, the computational approach to genome annotation. As it has been pointed by Stein [2001]: annotation is bridging the gap from sequence to the biology of the organism. All the steps required to improve the understanding of biological processes can be grouped into three categories to answer three complementary questions: where we can find the relevant information encoded in the sequence (the gene-level annotation); what roles the products of the gene expression play (the function-level annotation); and, how the genes and their products are integrated into a network of interactions (the process-level annotation).

In the late eighties, obtaining the genome sequence of a single eukaryotic organism, the human genome of course, was seen as a giant enterprise, that could only be tackled by an international consortium of research centers in a coordinated long term project. Although initially scheduled over fifteen years, as sequencing technology improved, faraway deadlines became closer, specially because of process automation. But it was the introduction of shotgun methodology what really spurred the production of huge eukaryotic genomes. The method heavily relies on the computational assembly of a myriad of sequenced fragments. It was first applied to produce bacterial genomes after which the team at Celera Genomics demonstrated its scalability to larger genomes by obtaining, in about a year, the genome sequence of *Drosophila melanogaster* [Adams *et al.*, 2000]. The competition between Celera and the Public Consortium yielded early results with the publication of the first draft version of the human genome in 2001 [Venter *et al.*, 2001; Lander *et al.*, 2001]. Nowadays, several large eukaryotic genome projects are undergoing, with a rate of one per year being published. The future will bring better sequences for more individuals and in less time. Examples of current developments for those forthcoming technologies were described by Kling [2003].

On the other hand, computational power has increased along with the availability of novel algorithms to analyze data. Traditional hypothesis testing is being more than complemented with the acquisition of large-scale data sets to which pattern recognition and data mining techniques are applied. The patterns arising from such analyses suggest novel hypotheses to test, while hypotheses can be tested directly using databases. Another milestone that must be taken into account is the development of the internet technologies during the last decade. The widespread use of the web to share data, software to analyze it and knowledge, has caused a revolution in science, among other subjects of our lives. It has also changed the way collaborative projects among groups all around the world can tackle larger and deeper analyses.

I have been part of this incessant flow of knowledge, of this never-ending endeavour, in which the analysis of genomes has become a key element. Writing this dissertation was like a stop in the road. Not only a break to rest, but also a time to think over, in order to gain an insight of what has been done, what is going on around and what can be done in the near future, before jumping again into the fast rivers of Genomics. In other words, I have tried to summarize my contribution to this field, grouping topics by their relationship rather than chronologically.

It is amazing how the availability of each new species genome can enhance our knowledge, not only of our own species, but also of life on Earth. I hope this grain of sand from the shores of Genomics will satisfy your scientific interest.

Josep Francesc Abril Ferrando
Barcelona, May 2005

Contents

Appendices

List of Tables

List of Figures

Acknowledgements

> Gratitude is born in hearts that take time to count up past mercies.
>
> —Charles E. Jefferson

I am grateful to my wife Marta for her constant support to continue the unpredictable endeavour that scientific research is. Since we met, she has been helping me more than she probably will imagine. I hope she will ever forgive me for my dedication to work and computers if I ever failed to give her attention. To her not only my deepest love but also my most grateful acknowledgements. Thank you for bringing into this world the cutest and most precious little girl I have ever met, my daughter Ruth. She also helped her father in her own way, just by being herself, making me happy, raising my spirit and, of course, encouraging me to keep going on those days you feel truly blue or stressed-out. Many thanks to my parents, for encouraging me to study when I was young, for sharing in the distance all our achievements, for their enthusiasm. Thanks also to my parents-in-law for all their support, for adopting me as a son, for the relaxing family Sunday dinners.

After several years working at a research institute it is not difficult to have met lots of interesting people, many of them really impressive. Therefore, I must first apologize for anybody who will think he or she has been left out. Those who know me, are already aware that it is easier for me to remember a face than a name. So that, I have tried to make up my mind and walk along my memories. I have to mention the long list of friends made in the Research Group in Biomedical Informatics (RGBI, or GRIB in Catalan). Not only for encouraging and helpful discussions; for the funny chitchats at coffee breaks; for sharing knowledge, code, data, and sometimes efforts too; for enjoying my jokes—although I have to admit that those were quite often uncomprehensible to the point that they were suffering rather than enjoying them—; for all the parties—and my apologies for attending less of them that I wanted to, thanks for understanding that I am a family man—.

I will begin with the old timers, they were already in the Research Group in Biomedical Informatics, when I started. They introduced me to *nix, to networks, to free software, to Bioinformatics. They also showed me what scientific research looks like. Juanjo Lozano, Moisès Burset and Jordi Rodrigo, I have to admit that it was a pleasure to meet you three, the most hilarious triplet I have ever seen. When I began, Roderic's team was only him and Moisès, and few undergraduate students—me, Jesus Feliu and David Alarcón—. I already met Jesus and David at the School of Biological Sciences of Universitat de Barcelona. We were part of a gang of computer maniacs that were regularly meeting to share tips and tricks, journals and programs. From the triplet, I was the only one who kept up doing research; so that it was a pleasing surprise when David Alarcón joined Baldo Oliva's group recently. Specially thanks to Juanjo and Moisès for struggling to install Linux in a machine despite the mistrust and arguments against such a system from the informatics support team of the center at that time.

I would like to thank Genís Parra, Sergi Castellano, Enrique Blanco, Charles Chapple, Nicolás Bellora, Francisco Câmara, Juan Antonio de los Cobos, Hugo Gutiérrez de Terán, Josep Pareja, Montserrat Barbany, Cristina Dezi, Fabien Fontaine, Elisabeth Gregori, Ramón Aragüés, Julio Bonis, Joan Planas, Adrián López, Alfons Nonell, Ruth Garriga, Jorge Naranjo, Lulla Opatowski, Cristina Herraiz, Pilar Noguerón, Claudio Silveira and Nuria Boada. Many, many thanks to Robert Castelo, Jan-Jaap Wesselink, Mar Albà, Eduardo Eyras, Jordi Villà, Baldomero Oliva, Nuria Centeno, Manolo Pastor, Jordi Mestres. Further thanks to Miguel Pignatelli, Alberto Roverato, Juan Valcárcel, Lluis Armengol, Mónica Bayés, Xavier Estivill, Marta Soldevila, Aida Andrés, Jordi Clarimón, Jaume Bertranpetit, Viviana Belalcázar and Marta Tomàs. I do not also forget those who visited us, Noura Dabbouseh, Marcos Rodrígues, Rachid Kara, Vanessa Adaui, David González, Juan Carlos Sánchez and Diego Miranda.

Of course, I have a special mention to our system administrators, Alfons González, Xavier Fustero and Òscar González. Not only because of friendship, but also because our work depends in great manner on their task and they are always patient with our endless requirements. Thanks for their helpful hints for solving this or that installation problem, sometimes related to my computer at home.

My deepest gratitude to those people from our group who helped me to review and proof-read this document. I would like here to point out and acknowledge the time spent, the comments, the corrections and suggestions made by Jan-Jaap Wesselink, Enrique Blanco, Charles Chapple, Òscar González and my wife, Marta, to this dissertation. Thanks again to Jan-Jaap for his commitment and his exhaustive proof-reading of this work. Further thanks to Robert Castelo for providing us the LaTeX files from his PhD thesis and for introducing us to `PDFlatex`. His templates were extended by Sergi Castellano and Genís Parra for their theses. The templates on which this document was built upon were derived from them.

To the secretaries that have been working for the group or for the IMIM along the time I have been there. Esther Román, Maite Cebrián, Yolanda Losada, Raquel Furió, Esther Callizo, Mireia Gusi, Nathalie Villahoz, and the veteran, Mercedes Fuertes. Thanks for their affection, for the chit-chats about our families, specially about our kids. To Eva Molero and Carlos Díaz. Further thanks to Alba Valls, Cristina García and Teresa Duran for their assistance in all the issues related with the PhD courses and, of course, the proceedings to submit and defend this thesis.

Thanks to the users of our software, especially those contributing with bug reports and/or patches to fix them, that interaction made those tools more useful. We appreciate their patience when the responsibilities of our own research took precedence over improving and maintaining the software. To those people who motivated and encouraged us to develop `gff2ps`, specially to Elena Casacuberta and Ampar Monfort. To Martin Reese, Sussana Lewis and Michael Ashburner, for allowing us to contribute to the GASP tutorial at ISMB99 meeting. The three-panel poster summarizing the results of the gene-prediction assessment were the first big dataset in which we tested `gff2ps`. Further thanks to Thomas Wiehe for initial suggestions for developing `gff2aplot` and latter involvement in its implementation; to Steffi Gebauer-Jung for providing parsers for alignment tools other than BLAST. To those people who motivated and encouraged us to continue improving it, specially to Matthias Plattzer; to those who gave valuable comments regarding this tool, as Web Miller.

I would like to thank Jim Fickett, for inviting Roderic and me, to SmithKline-Beecham (now Glaxo-Smithkline) research center in Philadelphia. It was my first trip to the States. There we met Pankaj Agarwal and I was able to see how a big pharmaceutical company looks like. To the people at Institut für Molekulare Biotechnologie (IMB), Jena; specially thanks to Matthias Plattzer, Gernot Glöckner, Karol Szafranski, Rüdiger Lehmann and Cornelia Baumgart. I wish to thank Thomas Wiehe and Steffi Gebauer-Jung for their friendliness and all the warm scientific collaborations with them, also for their hospitality when visiting them in Germany.

To the people at Celera Genomics at Rockville, Maryland, who got in contact with us to collaborate with the visualization of the fruit fly, the human and the mosquito genomes. Those collaborations allowed us to jump into the genomics field, moving from single gene sequences to work with whole genomes, from individual work to big collaborative efforts to solve one of the most complex problems to date. On the personal side, the warm welcome and all their attention, the opportunity to become part of such team of great minds, will be always in my heart. Thanks to Jennifer R. Wortman, Mark D. Adams, Patrick Dunn, Mark Yandell, William Majoros, Richard J. Mural, Robert A. Holt, George L. Gabor Miklos, Catherine Nelson, Gangadharan Subramanian (Mani), and J. Craig Venter. Thanks also to the *Drosophila melanogaster* jamboree people, specially to Gerald M. Rubin and Nomi L. Harris.

To the people at the international consortia for the sequencing and analysis of the mouse, rat and chicken genomes. For sharing preliminary data and knowledge, for the willingness in solving problems, for the endless conference calls, and so on. The list of people involved in such large projects is too big, but few people stand out by their exceptional organizational effort, such as Kim Worley, Victoria Hagigi and Ladeana Hillier. To Ewan Birney and Jim Kent, for ENSEMBL and GOLDEN PATH respectively, for replying to a mail as soon as it was sent, and for “wise” and funny discussions too. Further thanks to Web Miller, Peer Bork, Ivica Letunic, Chris Pontig, Donna Karolchik, Adam Siepel, David Haussler, Robert Baertsch, Ian Korf, Michael R. Brent, Chris Burge, Lior Pachter, Arian Smith, Emmanouil T. Dermitzakis, Alexandre Reymond, and Stylianos Antonarakis among others.

The publication of the first draft of the human genome had a tremendous impact on the media. We already had a small contact with journalists because of our participation in the fruit fly genome, reported just one year before. For the human genome that was not the case. Despite our small contribution, our group was the only Spanish partner directly involved in this huge project—unfortunately, boosting science in Spain was not one of the government priorities for a long time—. We were overwhelmed by interviews for newspapers and for radio and television programs. Elvira López and Maite Cebrián helped us to cope with them and to organize the appointments agenda for those “mad” weeks. This was when we met Marc Permanyer, from the Press department of Universitat Pompeu Fabra. The experience served, at least, to get more organized in advance, preparing press releases and concentrating interviews into press conferences. Thanks again to Elvira López, Marta Calsina and Marc Permanyer for organizing the press for the mouse, rat and chicken genomes. Further thanks to our group secretaries, for buffering all the incoming visits and telephone calls. Having cameras, photographers and journalists interfered with the work of many other members of our lab. Their patience and sense of humor must be acknowledged too. I would like to stress from this lines the contribution to the divulgation of scientific discovery in general, of our contributions in particular, made by many journalists. Among them, I would like to acknowledge Josep Corbella ("La Vanguardia"), Joaquim Elcacho ("Diari Avui"), Xavier Pujol Gebellí ("El Pais"), Antonio Madridejos ("El Periodico") and Javier López Rejas ("El Mundo").

To those great speakers that demonstrate their love for what they are doing, Roderic Guigó, Alfonso Valencia, Antonio Marín, Modesto Orozco and so on... I specially recall, not without fear but also with laughter, the de-construction of Bioinformatics lecture by Alfonso Valencia. On the other hand, I wish to thank all those who invited me to give talks about our work. To the Departmento de Biología y Geología of the Instituto de Enseñanza Secundaria Sanje at Alcantarilla, Murcia, for their warm welcome and for the interest demonstrated by students and teachers; specially thanks to Eva Palacios and Ángel Martínez. To Lola Andrade at the public library of Masnou, Barcelona, and the town council of Sant Carles de la Ràpita, Tarragona, particularly to Elvira Franquet i Tudó, Miguel Alonso Herrera and Josep Pere Geira, for inviting me to talk about the human genome too. I wish to thank also the organization committee of the meeting of the Sociedad Española de Genética, specially to José L. Oliver, for inviting me to present our research in their annual meeting held in El Escorial in 2003. I would like to thank the people of Fundació "La Caixa" for inviting us to organize the workshops on "Computational Analysis of DNA Sequences", held in Barcelona and Madrid; specially to Sílvia Maldonado, Sílvia Godó and Gloria Trías. Further thanks to Residència d'Investigadors for inviting us to the great première of *Verbum* ("Genoma in musica"), a piece for piano composed by Joan Guinjoan.

I also have many things to be thankful for to Ferran Sanz; for starting up the Research Group in Biomedical Informatics, which we have been part of; for his constant search of the excellence in science; for his wisdom and willingness to help, not only at scientific and academic levels; for his capacity to take new projects; for the Viladrau group retreats.

I am grateful to my PhD advisor, Roderic Guigó, for pushing us far beyond and at the same time for his patience when any analysis took longer than expected—and for his famous sentence, "that would take just five minutes, wouldn't it?"—; for his insights in the field and his dedication to science—I agree that science is not just a job but a way of life, although one needs to make a living too—; for introducing me to `gawk`, POSTSCRIPT and LaTeX. For his efforts to get funds for the research, as we all know how much it takes to fill in all the bureaucracy related to a project—and how this interferes with the "field" work. For introducing us to outstanding people in the field. For his deadline last minute questions. For a scholarship in 1998/1999, that permitted to devote myself full-time to research. For a scholarship in 2003/2004, mainly because that allowed me to finalize my PhD thesis and write this dissertation.

I acknowledge the support from the Instituto de Salud Carlos III (ISCIII) for the Beca de Formación de personal Investigador (BEFI, a PhD studentship) for the 1999/2003 four years period. Thanks to the ISMB'99 organizing committee for a travel scholarship to attend to the 1999 meeting held in Heidelberg (Germany). To a joint grant from the German Academic Exchange Service (DAAD) to Thomas Wiehe and the Ministerio de Educación y Ciencia (Spain) to Roderic Guigó, which made possible, among other things, our scientific stays in the Institut für Molekulare Biotechnologie (IMB) at Jena (Germany). I would also mention all the people that, from the Federación de Jóvenes Investigadores and the Precarios association, are trying to improve the labor situation of research scholarships. Thanks to their efforts current and future generations of PhD students will have hopefully better work conditions than us. Special thanks to Sergi Castellano for getting so involved with Precarios and getting us up to date with their activities and achievements.

Finally, I would like to thank Linus Torwalds (for developing the Linux kernel), the GNU Free-Software Foundation (for `bash`, `gawk`, `make` and a myriad of other useful *NIX tools, but also for the advocacy of free software), Larry Wall *et al* (for the `perl` programming language, its dynamic community and the useful modules from CPAN), Richard M. Stallman *et al* (for the arguably best programming text editor, `emacs`, and my apologies to `vi` advocators), Norman Ramsey (for the `noweb` literate programming tool), Donald Knuth and Lesslie Lamport (for the TeX and LaTeX typesetting systems, respectively—and thanks to all the developers of a whole bunch of useful packages, such as `PStricks`, `PDFLaTeX`, `natbib`, `makeidx`, `hyperref`, and so on—), Sergie Brin and Larry Page (for devising the page rank technology behind `Google`). To the efforts of all the people that have demonstrated that sharing will hopefully provide a better future. All the advances of humankind are the result of the accumulation of knowledge. All their contributions provided us with the tools with which we have performed our analyses and developed our software.

To all of you, many, many thanks from the heart…

Abstract

The constantly increasing amount of available genome sequences, along with an increasing number of experimental techniques, will help to produce the complete catalog of cellular functions for different organisms, including humans. Such a catalog will define the base from which we will better understand how organisms work at the molecular level. At the same time it will shed light on which changes are associated with disease. Therefore, the raw sequence from genome sequencing projects is worthless without the complete analysis and further annotation of the genomic features that define those functions. This dissertation presents our contribution to three related aspects of gene annotation on eukaryotic genomes.

First, a comparison at sequence level of human and mouse genomes was performed by developing a semi-automatic analysis pipeline. The SGP2 gene-finding tool was developed from procedures used in this pipeline. The concept behind SGP2 is that similarity regions obtained by TBLASTX are used to increase the score of exons predicted by geneid, in order to produce a more accurate set of gene structures. SGP2 provides a specificity that is high enough for its predictions to be experimentally verified by RT-PCR. The RT-PCR validation of predicted splice junctions also serves as example of how combined computational and experimental approaches will yield the best results.

Then, we performed a descriptive analysis at sequence level of the splice site signals from a reliable set of orthologous genes for human, mouse, rat and chicken. We have explored the differences at nucleotide sequence level between U2 and U12 for the set of orthologous introns derived from those genes. We found that orthologous splice signals between human and rodents and within rodents are more conserved than unrelated splice sites. However, additional conservation can be explained mostly by background intron conservation. Additional conservation over background is detectable in orthologous mammalian and chicken splice sites. Our results also indicate that the U2 and U12 intron classes have evolved independently since the split of mammals and birds. We found neither convincing case of interconversion between these two classes in our sets of orthologous introns, nor any single case of switching between AT-AC and GT-AG subtypes within U12 introns. In contrast, switching between GT-AG and GC-AG U2 subtypes does not appear to be unusual.

Finally, we implemented visualization tools to integrate annotation features for gene-finding and comparative analyses. One of those tools, gff2ps, was used to draw the whole genome maps for human, fruitfly and mosquito. gff2aplot and the accompanying parsers facilitate the task of integrating sequence annotations with the output of homology-based tools, like BLAST. We have also adapted the concept of pictograms to the comparative analysis of orthologous splice sites, by developing compi.

Resum

L'incessant augment del nombre de seqüències genòmiques, juntament amb l'increment del nombre de tècniques experimentals de les que es disposa, permetrà obtenir el catàleg complet de les funcions cel·lulars de diferents organismes, incloent-hi la nostra espècie. Aquest catàleg definirà els fonaments sobre els que es podrà entendre millor com els organismes funcionen a nivell molecular. Al mateix temps es tindran més pistes sobre els canvis que estan associats amb les malalties. Per tant, la seqüència en brut, tal i com s'obté dels projectes de seqüenciació de genomes, no té cap valor sense les anàlisis i la subsegüent anotació de les característiques que defineixen aquestes funcions. Aquesta tesi presenta la nostra contribució en tres aspectes relacionats de l'anotació dels gens en genomes eucariotes.

Primer, la comparació a nivell de seqüència entre els genomes humà i de ratolí es va dur a terme mitjançant un protocol semi-automàtic. El programa de predicció de gens SGP2 es va desenvolupar a partir d'elements d'aquest protocol. El concepte al darrera de l'SGP2 és que les regions de similaritat obtingudes amb el programa TBLASTX, es fan servir per augmentar la puntuació dels exons predits pel programa geneid, amb el que s'obtenen conjunts d'anotacions més acurats d'estructures gèniques. SGP2 té una especificitat que és prou gran com per que es puguin validar experimentalment via RT-PCR. La validació de llocs d'*splicing* emprant la tècnica de la RT-PCR és un bon exemple de com la combinació d'aproximacions computacionals i experimentals produeix millors resultats que per separat.

S'ha dut a terme l'anàlisi descriptiva a nivell de seqüència dels llocs d'*splicing* obtinguts sobre un conjunt fiable de gens ortòlegs per humà, ratolí, rata i pollastre. S'han explorat les diferències a nivell de nucleòtid entre llocs U2 i U12, pel conjunt d'introns ortòlegs que se'n deriva d'aquests gens. S'ha trobat que els senyals d'*splicing* ortòlegs entre humà i rossegadors, així com entre rossegadors, estan més conservats que els llocs no relacionats. Aquesta conservació addicional pot ser explicada però a nivell de conservació basal dels introns. D'altra banda, s'ha detectat més conservació de l'esperada entre llocs d'*splicing* ortòlegs entre mamífers i pollastre. Els resultats obtinguts també indiquen que les classes intròniques U2 i U12 han evolucionat independentment des de l'ancestre comú dels mamífers i les aus. Tampoc s'ha trobat cap cas convincent d'interconversió entre aquestes dues classes en el conjunt d'introns ortòlegs generat, ni cap cas de substitució entre els subtipus AT-AC i GT-AG d'introns U12. Al contrari, el pas de GT-AG a GC-AG, i viceversa, en introns U2 no sembla ser inusual.

Finalment, s'han implementat una sèrie d'eines de visualització per integrar anotacions obtingudes pels programes de predicció de gens i per les anàlisis comparatives sobre genomes. Una d'aquestes eines, el gff2ps, s'ha emprat en la cartografia dels genomes humà, de la mosca del vinagre i del mosquit de la malària, entre d'altres. El programa gff2aplot i els filtres associats, han facilitat la tasca d'integrar anotacions de seqüència amb els resultats d'eines per la cerca d'homologia, com ara el BLAST. S'ha adaptat també el concepte de pictograma a l'anàlisi comparativa de llocs d'*splicing* ortòlegs, amb el desenvolupament del programa compi.

Resumen

El aumento incesante del número de secuencias genómicas, junto con el incremento del número de técnicas experimentales de las que se dispone, permitirá la obtención del catálogo completo de las funciones celulares de los diferentes organismos, incluida nuestra especie. Este catálogo definirá las bases sobre las que se pueda entender mejor el funcionamiento de los organismos a nivel molecular. Al mismo tiempo, se obtendrán más pistas sobre los cambios asociados a enfermedades. Por tanto, la secuencia en bruto, tal y como se obtiene en los proyectos de secuenciación masiva, no tiene ningún valor sin los análisis y la posterior anotación de las características que definen estas funciones. Esta tesis presenta nuestra contribución a tres aspectos relacionados de la anotación de los genes en genomas eucariotas.

Primero, la comparación a nivel de secuencia entre el genoma humano y el de ratón se llevó a cabo mediante un protocolo semi-automático. El programa de predicción de genes SGP2 se desarrolló a partir de elementos de dicho protocolo. El concepto sobre el que se fundamenta el SGP2 es que las regiones de similaridad obtenidas con el programa TBLASTX, se utilizan para aumentar la puntuación de los exones predichos por el programa geneid, con lo que se obtienen conjuntos más precisos de anotaciones de estructuras génicas. SGP2 tiene una especificidad suficiente como para validar esas anotaciones experimentalmente vía RT-PCR. La validación de los sitios de *splicing* mediante el uso de la técnica de la RT-PCR es un buen ejemplo de cómo la combinación de aproximaciones computacionales y experimentales produce mejores resultados que por separado.

Se ha llevado a cabo el análisis descriptivo a nivel de secuencia de los sitios de *splicing* obtenidos sobre un conjunto fiable de genes ortólogos para humano, ratón, rata y pollo. Se han explorado las diferencias a nivel de nucleótido entre sitios U2 y U12 para el conjunto de intrones ortólogos derivado de esos genes. Se ha visto que las señales de *splicing* ortólogas entre humanos y roedores, así como entre roedores, están más conservadas que las no ortólogas. Esta conservación puede ser explicada en parte a nivel de conservación basal de los intrones. Por otro lado, se ha detectado mayor conservación de la esperada entre sitios de *splicing* ortólogos entre mamíferos y pollo. Los resultados obtenidos indican también que las clases intrónicas U2 y U12 han evolucionado independientemente desde el ancestro común de mamíferos y aves. Tampoco se ha hallado ningún caso convincente de interconversión entre estas dos clases en el conjunto de intrones ortólogos generado, ni ningún caso de substitución entre los subtipos AT-AC y GT-AG en intrones U12. Por el contrario, el paso de GT-AG a GC-AG, y viceversa, en intrones U2 no parece ser inusual.

Finalmente, se han implementado una serie de herramientas de visualización para integrar anotaciones obtenidas por los programas de predicción de genes y por los análisis comparativos sobre genomas. Una de estas herramientas, gff2ps, se ha utilizado para cartografiar los genomas humano, de la mosca del vinagre y del mosquito de la malaria. El programa gff2aplot y los filtros asociados, han facilitado la tarea de integrar anotaciones a nivel de secuencia con los resultados obtenidos por herramientas de búsqueda de homología, como BLAST. Se ha adaptado también el concepto de pictograma al análisis comparativo de los sitios de *splicing* ortólogos, con el desarrollo del programa compi.

Chapter 1

Introduction

> All our progress is an unfolding, like vegetable bud. You have first an instinct, then an opinion, then a knowledge
>
> —Ralph Waldo Emerson, "*Essays*"

Genes encode all the information necessary for the cell to carry out all its functions. Although protein sequences are continuous[1], the sequence of the genes defining them in the eukaryotic organisms appears in the DNA sequence interspersed in a sea of non-coding regions. Furthermore, evolution has made the problem of finding those genes in anonymous DNA sequences harder. Not only because of the intrinsic mutational changes of the DNA sequences, which makes homology finding more difficult; but also due to the variation accumulated in the gene catalog of each species, which has been expanded—by duplications, for instance— or reduced—i.e., by deletions and lose of function (pseudogenes). In addition to that, genes have been reordered, some of them have lost their function, becoming useless, and so on. On the other hand, to search for genes means that we have to look for the features that characterize them, examining the raw DNA sequences for the signals that delineate them. Therefore, obtaining the genome sequence of an organism does not grant that we will be able to find all the genes easily, as the real ones will be hidden in a forest of false signals and real non-coding regions. The fact that in the human genome, made up of three billions[2] of nucleic acids distributed in 23 chromosomes (the haploid set of course), there is only about 2% of sequence in coding regions, helps us to understand the magnitude of the problem of finding the genes encoded in it [Guigó *et al*., 2000; Venter *et al*., 2001; Lander *et al*., 2001].

At the moment of transcription, the sequence containing a gene is copied from the DNA to RNA, the so called primary transcript. This undergoes a series of modifications before being transported from the nucleus to the cytoplasm. Once there, the sequence of the RNA, known at this step as messenger RNA (mRNA), serves as a template to produce the corresponding protein, the translation process. The pathway from DNA to protein synthesis became the central dogma of Biology. One of the most important changes performed on

[1] Genes that do not translate into proteins can still have a function, such as the transfer RNA (tRNA) genes and other non-coding RNAs (ncRNA). Whatever they are still coding for a cellular function, the term coding will be used along this document as protein-coding, as for protein-coding genes.

[2] US notation: 3×10^9, more intuitively 3,000,000,000 bp.

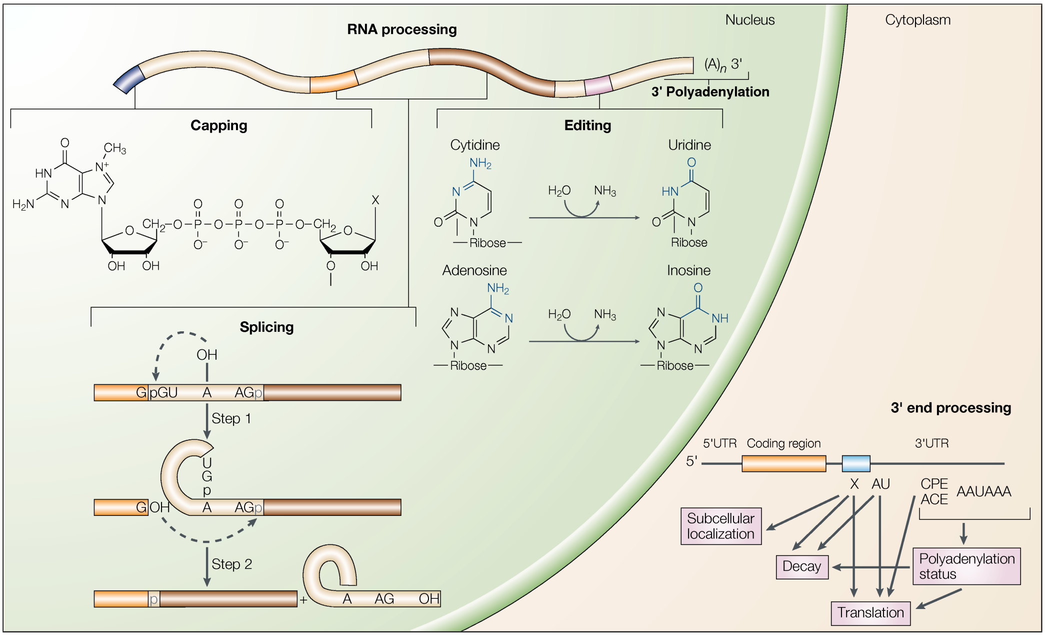


Figure 1.1: **The processing of RNA in the cell.** Immediately after the RNA is transcribed in the nucleus, capping, splicing, editing and 3′ polyadenylation of the pre-mRNA occur. In mammals, RNA editing can be of two types, either the conversion of cytidine to uridine or the conversion of adenosine to inosine. Once the mRNA is transported into the cytoplasm, additional processing of the polyA tail can occur. The elements required for this and for subcellular localization, stability and translation are present in the untranslated regions (UTRs). Adapted from Keegan *et al.* [2001].

the primary transcript is the elimination of the fragments not coding for proteins, the so called introns, by means of a set of biochemical reactions in the cell nucleus, known as the splicing process. The final product of splicing is a molecule of mRNA in which the gene's exons have been concatenated to get a continuous gene sequence. Figure 1.1 illustrates the modifications that the primary transcript undergoes. Capping of the 5′ terminus, splicing of the exonic segments and polyadenylation are the major events leading to the mature mRNA molecule. All those steps can be coupled in the cell as has been suggested in recent publications [Proudfoot *et al.*, 2002; Zorio and Bentley, 2004].

The next challenge is how to delineate the exonic structures that define a gene product. Unlike prokaryotic organisms, for which genes are formed by a single exon—and the intergenic sequences, if present, are very short—, the eukaryotic genes can have more than one, up to hundreds in some cases. In the human genome, for example, approximately a 10% of the 33,000 genes annotated in the last human genome version[3] are single exon genes, and all the rest are multi-exonic gene structures. The following big problem, yet to be solved, is to find all the alternative exonic structures encoded in a given gene region, what is also known as alternative splicing . Recent estimates suggest that more than 60% of human

[3] Calculated from ENSEMBL genes found in the GOLDEN PATH HG16 version (July, 2003), obtained from: http://hgdownload.cse.ucsc.edu/goldenPath/hg16/database/ensGene.txt.gz

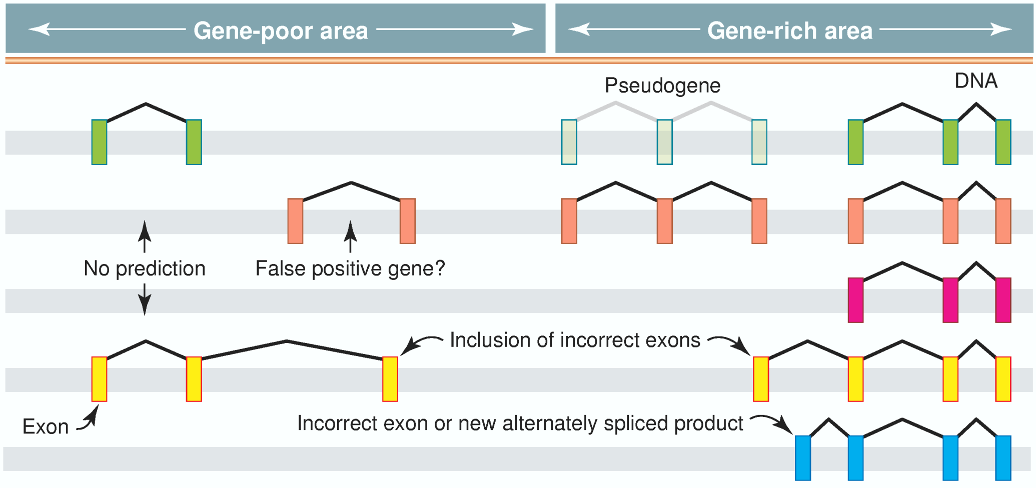


Figure 1.2: **Common pitfalls among gene-finding approaches.** No program is yet able to find all genes in anonymous genomic sequences correctly. Some overpredict and report genes where there are none; some misspredict genes; in other cases they are not able to properly group exons belonging to one or more genes, joining or splitting the corresponding gene structures. The upper track shows a putative set of real genes, the other tracks simulate the output of four different gene-finding tools. Adapted from Pennisi [2003].

genes show this phenomenon [Lander *et al.*, 2001; Modrek *et al.*, 2001]. Landscape becomes more complex if one wants to take into account the regulation of gene expression [Zhang, 2002] and the rules of the alternative splicing control [Woodley and Valcárcel, 2002].

1.1 Finding Genes in the Genomes

In the early eighties, DNA sequences under analysis were long enough to find initially open reading frames (ORFs), then exons. The first computational approaches focused then on the search for coding regions—see, for example, the pioneering works of Pustell and Kafatos [1982], Staden [`ANALYSEQ` and the `Staden` package, 1984b; 1986 respectively], Devereux *et al.* [`GCG` suite, 1984], Keller *et al.* [1984], or Blattner and Schroeder [1984]. It was not until the nineties that programs able to assemble those exons into a complete gene were developed [Uberbacher and Mural, 1991; Guigó *et al.*, 1992; Burge and Karlin, 1997]. Although sequencing technology was improving, most of the available sequences contained a single gene, often incomplete. By that time, the number of sequences stored in databases was relatively small. Whole genome sequencing projects changed that scenario. Databases started to grow exponentially and new problems had to be faced by the sequence analysis algorithms. Speed was one of the main requirements of the new era, not only to look for genes but also for the search of homologies between sequences of different species, mapping repetitive sequences, and so on. Novel algorithms for homology search, less sensitive but faster, were developed to screen an ever growing set of sequences. Models underlying the gene-finding software were developed from different approaches—for instance, neural

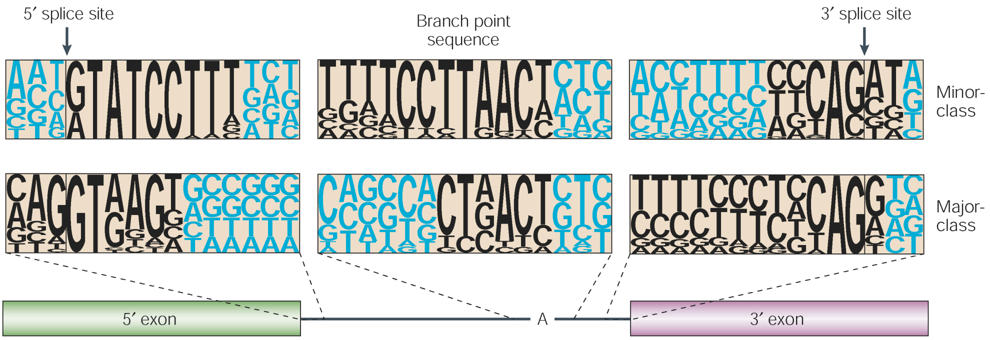


Figure 1.3: **Consensus sequences of U2 and U12 splicing signals.** The consensus sequences of the 5′ splice site, branch site and 3′ splice site are shown from left to right for minor-class introns (upper row) and for major-class introns (lower row). The letter heights at each position represent the frequency of occurrence of the corresponding nucleotides at that position. The positions that are thought to be involved in intron recognition are shown in black; other positions are shown in blue. Adapted from Patel and Steitz [2003].

networks and hidden Markov models (HMMs). However, as the length of the sequences increased, it was evident that gene distribution along them and their structural complexity became a hard problem to solve. The reliability of the results obtained by computational gene prediction tools has not improved so fast [Burset and Guigó, 1996; Guigó *et al.*, 2000; Reese *et al.*, 2000].

Gene prediction has changed substantially in the past few years. The sequencing of an increasing number of eukaryotic genomes, and the distribution through centralized genome browsers,—such as those at the University of California Santa Cruz (UCSC), the National Center for Biotechnology Information (NCBI) and the European Bioinformatics Institute (EBI)—of precomputed genome-wide annotations may often make it unnecessary for scientists to run gene prediction programs themselves. Gene prediction, however, is still useful in these genomes, because researchers may want, for instance, to investigate in detail the pattern of alternative splicing of a given gene. On the other hand, gene prediction is still essential to analyze sequences from the many genomes that have not been completely characterized yet. The obvious conclusion is that gene prediction is still an open problem. Figure 1.2 highlights some of the common failings that the current tools have yet to overcome.

Chapter 3 presents a brief overview of gene finding, both classical and comparative approaches, and the evaluation of the predictions, as well as a description of the semi-automatic protocols used for large genome-sized data sets.

1.2 Eukaryotic Gene Structure

The precise removal of pre-mRNA introns is a critical aspect of gene expression. The splicing machinery must recognize and remove introns to make the correct message for protein production, but also, for many genes, alternative splicing mechanisms must be in place to generate functionally diverse protein isoforms in a spatially and temporally regulated manner [Hastings and Krainer, 2001]. Paradoxically, in higher eukaryotes, the requirement for accurate splicing is accompanied by exon-intron junctions that are defined, in most cases, by weakly conserved intronic *cis*-elements, the splice sites and the branch point [Cartegni *et al.*, 2002]. These elements are necessary but by no means sufficient to define exon-intron boundaries. Sequences that match the consensus splice site signals as well as, or better than, natural splice sites are very common in introns. They define a set of pseudo-exons that greatly outnumber genuine exons and greatly complicate the task of assembling real gene structures by the computational gene-finding approaches.

The splicing reaction is mediated by two distinct yet analogous pools of small nuclear ribonucleoprotein particles. The RNA component of such particles takes part in the recognition of sequence motifs at both ends of the introns, the $5'$ and $3'$ splice sites, and a region within the intron known as the branch point [Patel and Steitz, 2003]. The works of Hall and Padgett [1994] revealed a minor class of introns having unusual consensus splice site sequences. Figure 1.3 shows, side by side, the sequence patterns for both the major and minor intron classes and illustrates the fact that the minor-class sequence motifs are far more conserved than those for the major-class [Sharp and Burge, 1997].

After a detailed description of the splicing biochemistry, we will focus on the sequence features that define the boundaries between exons and introns in chapter 4. Our contribution to understanding the biological characteristics of such features, based on the comparative analysis of introns from orthologous genes of several vertebrate genomes, is also described.

1.3 Visualizing Genomic Features

Despite substantial progress in computational gene finding, currently available methods are not yet able to automatically provide accurate enough descriptions of the gene content of eukaryotic genomes and a substantial amount of manual curation is required. This is a task in which visualization and integration tools play an essential role.

Any result in Bioinformatics, whether it is a sequence alignment, a structure prediction, or an analysis of gene expression patterns, should answer a biological question. For this reason, it is up to the researchers to interpret their results in the context of such a question. This interpretation is the most important part of the scientific process and a number of programs are used to visualize the sort of data arising from Bioinformatics research. These programs range from general-purpose plotting and statistical packages for the analysis of numerical data to programs dedicated to presenting sequence annotations in an integrated, intuitive and comprehensive fashion, such as the ENSEMBL genome browser examples from Figure 1.4. Visualization tools exploit the abilities of the eye and brain to find patterns that may be interesting. After that, statistical and data mining tools restrict those searches to the patterns that can be quantitatively and repeatedly shown to be significant [Gybas and Jambeck, 2003].

In chapter 5, we provide an overview of visualization tools that have been applied to the analysis of genome annotations and the inter-specific comparative analyses. Furthermore, we show a set of tools we have developed to visualize genomic annotations.

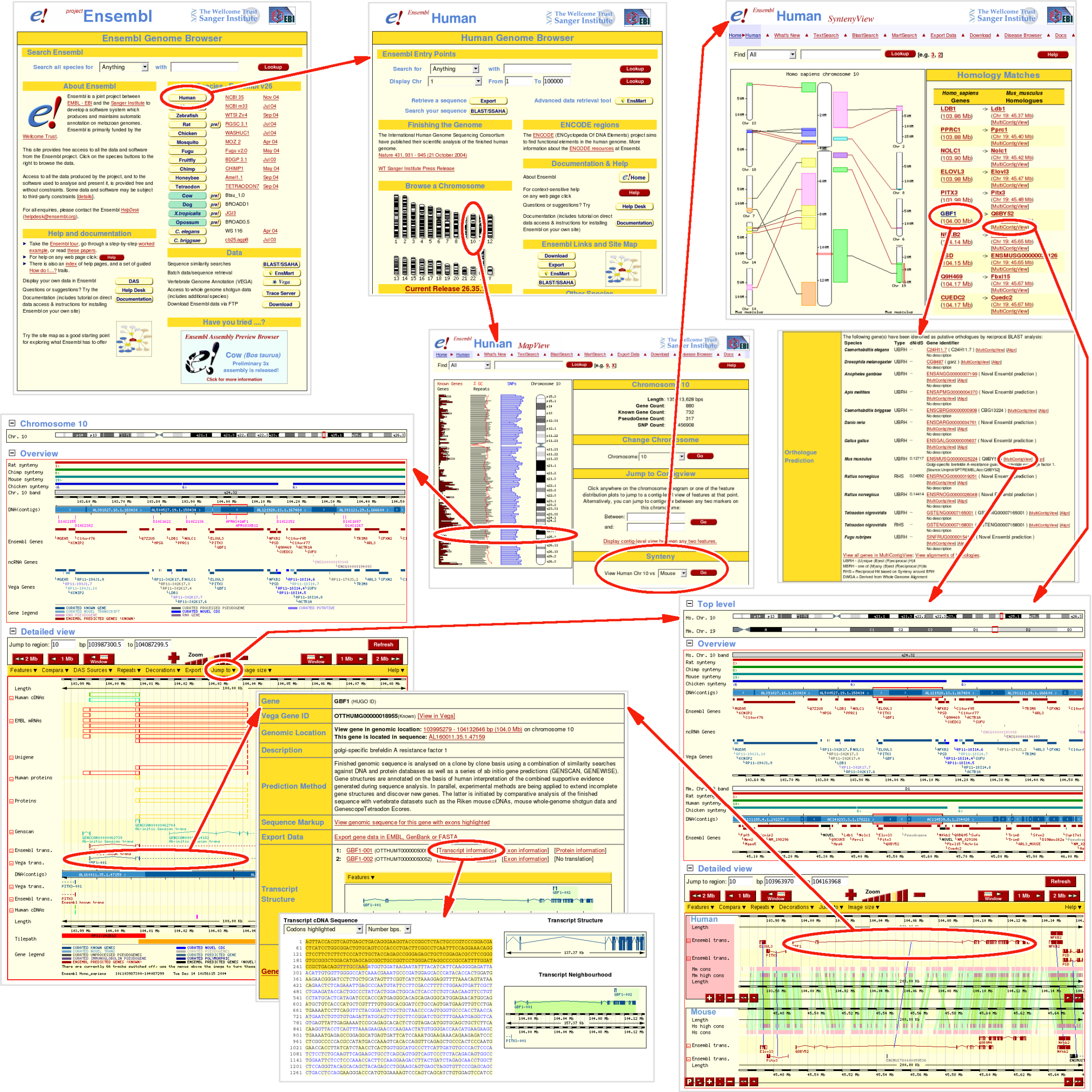


Figure 1.4: **Browsing through genome annotations.** A quick tour through the ENSEMBL genome browser pinpoints the different information levels we can access via its web interface. From their home page located in the upper left corner, a researcher can jump into the desired genome, the human genome in this example. Specific queries can be performed by using the text forms, but a very intuitive interface allows the user to zoom from the chromosome level (the Map View window placed in the center of this figure), to the sequence level (the Contig Viewer on the lower left panel), and to the gene or transcript reports (middle lower panels). Integration with other species-specific genome databases is also possible by using the Synteny panels (upper right panel). Comparative analyses at the genomic sequence level are shown in the Multi-Contig View (lower right panel). Red arrows indicate only few of the possible paths a researcher can follow through this browser.

1.4 About This Thesis

None of the articles composing this thesis were collected in an appendix or as separate chapters. They appear as sections where links to the journal web references and supplementary material are provided, followed by the article itself. Presenting the publications this way may break the storyline but it puts related subjects together which seems to be more appropriate. In those papers in which we were part of an international consortium, the article is reproduced in part due to its size but also because we have atempted to focus on our specific contribution. This should not be a problem, since the link to retrieve the whole article is provided as was already mentioned. Several figures and tables are referred to along the text via hyperlinks pointing directly to the page of the corresponding embedded article. Absolute page numbers relative to this document were used in all of these hyperlinks and in the list of figures or tables. Nevertheless, the reader can find easily the original paper page numbers just by following the hyperlinks.

The electronic version of this document has hyperlinks for the table of contents, for the bibliographic references, but most important of all, also for the web addresses on the Internet—from now on, their Uniform Resource Locator (URL). This means that you can visit the corresponding web page by clicking your pointer on them, in case that you have your PDF viewer properly customized. Many of the URLs presented in this book have been collected in a web links reference index available on page 213. URLs within paragraphs have been moved into that web glossary in order to avoid unbalanced line breaks and for a more pleasant reading. A reference to the corresponding page in the web reference index is provided instead. That does not include those URLs refering to the supplementary materials of the attached articles, which are put together in the corresponding article section (see Section 3.2.1 in page 20 for an example).

An attempt has been made to keep software names as provided by their authors. Those names appear in a `monospaced serif font`. Database names are typeset in a SMALL-CAPS SANS-SERIF FONT. A *slanted sans-serif font* was used for gene names, while a upright sans-serif font was chosen for protein names.

The first time an acronym appears in the document, the full name will be provided and the acronym itself will be shown in parentheses. From then on, the short form will be used. In order to help the reader, a list of abbreviations can be found on page 203. A glossary of terms is also available on page 207.

Chapter 2
Objectives

Don't bite my finger, look where it's pointing.

—Warren S. McCulloch

The research in this *PhD thesis* was initially targeted, in late 1998, to the goals enumerated below. In what follows, they are described and an account of their achievement status given.

1. To analyze through bioinformatic means the exonic structures of homologous genes, in order to determine the extent of conservation at gene structure level.

2. To describe possible evolutive patterns for those exonic structures within mammals and vertebrates.

3. To compare the conservation of the signals that delineate exons between different species. Both, acceptor and donor, splice sites are the main players in the definition of the exonic structure of eukaryotic genes.

4. To investigate the relationship between the conservation of exonic structures and alternative splicing patterns.

5. To develop visualization tools focusing specifically on the annotation of genomic sequences (including output from gene finding tools) and the comparative analysis of exonic structures.

6. To provide and distribute the results of our analyses and the bioinformatic tools to the research community.

These objectives were established based on data and knowledge of that time. They were intended to explore very basic questions about the exonic structure of eukaryotic genes and the evolutionary fates of introns. These goals have been accomplished to different degrees as related further down. Therefore, several of these points should be considered as ongoing work and yet many questions, both old and new, remain unanswered.

Some of the work presented in this dissertation has been done in collaboration with international genome sequencing consortia. These collaborations gave me the opportunity to meet and work with specialists from all over the world, and made our work very relevant. However, those collaborations put a lot of pressure on us and a lot of effort has been invested in such genome annotation projects. On the other hand, participating in the annotation of recently sequenced genomes has proven fruitful, as we have had to develop methodologies to analyze large amounts of data from different sources for each species. This means that we had to implement specific software to solve new problems, as well as to establish protocols to handle large sequence and annotation data sets. Such an effort was detrimental to some of the initial objectives and it made that this thesis took more time than expected.

The protocols and software we developed for finding genes by the comparison of the human and mouse genomes [Parra *et al.*, 2003; Waterston *et al.*, 2002], have been adapted to produce gene annotations in a semi-automatic pipeline for each novel assembly version of eukaryotic genomes. Annotations for several species, including human, chimpanzee, mouse, rat, chicken and the fruitfly, are available through a web repository (see page 214, on Web Glossary).

Despite the fact that we were able to undertake the analysis of the orthologous splice sites for four vertebrate species, we have not been able to investigate the conservation of exonic structures of alternatively spliced isoforms of orthologous genes. We could not tackle the evolutionary analysis of exonic gene structure either. However, during the last year, our group has joined the Alternative Splicing Database Project [Thanaraj *et al.*, 2004], and has been also chosen as a partner of the ENCODE project [ENCODE Project Consortium, 2004]. ASD aims to analyze the mechanism of splicing on a genome-wide scale by creating both, human-curated and computer-generated databases containing alternatively spliced exons from human and other model species. The main aims of the ENCODE project are both to validate known genes and to confirm reliable computational predictions experimentally. However, also to identify previously unknown genes and the characterization of a number of splice variants of the genes found in the corresponding target regions. In both projects, there are people in our laboratory that will continue this promising research line.

For the last objective, all the programs and data sets have been made available through our group's web server. Most of our published papers have their own web page with supplementary materials, as can be seen in the corresponding sections. Regarding the visualization software developed, `gff2ps` and `gff2aplot`, both have several tutorials and a user's reference manual. Furthermore, these tools are distributed under the GNU General Public License (GNU-GPL). The GNU-GPL is intended to guarantee the freedom to share and change free software–to make sure the software is free for all its users. If our research is publicly funded, the fruits of our work should be made publicly available. Both, the GNU-GPL and the Internet, are in our honest opinion most forthright approach to accomplish that responsibility with the society. As stated in Jamison [2003], software security measures which don't allow for examination of original code or for reasonable mechanisms of validity testing are in contrast with the open communication needed to do science properly.

Chapter 3

Comparative Gene Finding

> When this circuit learns your job,
> what are you going to do ?
>
> —Herbert Marshall McLuhan

Life processes, from the information flow from DNA to proteins to biochemical or regulatory pathways, have an intrinsic algorithmic nature. An algorithm can be defined as a detailed sequence of actions to perform to accomplish some task. The cells of living beings steadily perform step-by-step chemical reactions. Interactions between molecules modulate the flow of energy or information across the cell. The analogy works the other way around, as we attempt to emulate such biological processes by computational methods. The organization of a gene, as any other biological structure, is determined by functional and evolutionary constraints. All computational methods are therefore based on our experimental understanding of such constraints.

In this chapter we explore the computational modeling of protein-coding gene structures. After that, we describe our contribution to the gene-finding using comparative genomics approaches.

3.1 Computational Gene Prediction

After the genome of an organism is sequenced and assembled, comprehensive and accurate initial gene prediction and annotation by computational analysis have become the necessary first step towards understanding the functional content of the genome [Guigó and Zhang, 2004]. Despite the fact that, in practice, there are tools that can be classified in more than one of them, we can split the computational approaches to find genes in DNA sequences into three main categories.

- "*Ab initio*" methods are based on a search for those signals that specify the boundaries of coding regions, as in the analysis of coding biases and regularities of the protein-coding versus non-coding regions [Guigó, 1999]. The main handicap of such approaches is that the molecular mechanisms used by eukaryotic cells to define the signals that determine the gene structure are not completely well understood.

- Homology-based methods use information related to the similarity of the query coding region with respect to a set of known sequences from databases. The major drawback here is the bias towards known genes or proteins. Therefore novel families that are under-represented or not found in the databases, will still be hard to retrieve [Guigó *et al.*, 2000].

- The whole-genome sequencing projects allowed to extend the previous approach. Instead of searching for sequences of known genes, the entire genomes of two or more species are compared. The idea behind this is that evolution tends to retain those regions that are important because they have a function, whatever it encodes: a protein or structural or regulatory elements. When comparing genomes of closely related species, a set of genes emerges that is characteristic for the taxonomic group to which they belong. A good example of this has been the comparison between the human [Lander *et al.*, 2001] and mouse [Waterston *et al.*, 2002] genomes, during which approximately 9,000 novel mouse and 1,000 novel human genes have been annotated [Guigó *et al.*, 2003; Flicek *et al.*, 2003; Parra *et al.*, 2003]. However, comparative genomics approaches are not only a useful tool to find novel genes, but they are also a tool to improve the annotations of known genes [Reichwald *et al.*, 2000] and to hypothesize about their functions [Wiehe *et al.*, 2000].

3.1.1 "*Ab initio*" developments

Computational gene finding is not a brand new field and a large body of literature has accumulated during the last 25 years. Early studies by Shepherd [1981], Fickett [1982] and Staden and McLachlan [1982] showed that statistical measures related to biases in amino acid and codon usage could be used to approximately identify protein coding regions in genomic sequences. Based on these differences, the first generation of gene predictions programs, designed to identify approximate locations of coding regions in genomic DNA, was developed. The most widely known of this kind of programs were probably `testcode` (based on Fickett [1982]) and `grail` [Uberbacher and Mural, 1991]. These programs were able to identify coding regions of sufficient length (100-200 bp) with fairly high reliability, but did not accurately predict exon locations.

In order to predict exon boundaries, a new generation of algorithms was developed. A second generation of programs, such as `sorfind` [Hutchinson and Hayden, 1992], `grailII` [Xu *et al.*, 1994b,a] and `xpound` [Thomas and Skolnick, 1994], uses a combination of splice signal and coding region identification techniques to predict potential sets of exons (spliceable open reading frames), but does not attempt to assemble predicted exons into complete genes. A third generation of programs attempts the more difficult task of predicting complete gene structures: sets of exons which can be assembled into translatable coding sequences. The earliest examples of such integrated gene finding algorithms were probably the `genemodeler` program [Fields and Soderlund, 1990] for prediction of genes in *Caenorhabditis elegans* and the method of Gelfand [1990] for mammalian sequences. Subsequently, there has been a mini-boom of interest in development of such methods and a wide variety of programs have appeared, including: `geneid` [Guigó *et al.*, 1992], which used a hierarchical rule-based structure; `geneparser` [Snyder and Stormo, 1993], which scored all subintervals in a sequence for content statistics and splice site signals, then weighted them by a neural network and it chained the resulting features by dynamic programing; genemark [Borodovsky and McIninch, 1993] which combined the specific Markov models of coding and non-coding region together with Bayes' decision making function; genlang [Dong and Searls, 1994], which treated the problem by linguistic methods describing a grammar and parser for eukaryotic protein-encoding genes; and fgenes [Solovyev *et al.*, 1994] which used a discriminant analysis for identification of splice sites, exons and promoter elements.

At the end of the last decade, the introduction of the Generalized Hidden Markov Models (GHMMs) produced a new generation of gene prediction programs. GHMMs have some advantages over the previous approaches. The main advantage is that all the parameters of the model are probabilities and that, given a set of curated sequences and defined states, the Viterbi algorithm can be used to compute the set of optimal parameters. A great variety of programs appeared simultaneously exploring the capabilities of GHMMs: genie [Kulp *et al.*, 1996], hmmgene [Krogh, 1997], veil [Henderson *et al.*, 1997], genscan [Burge and Karlin, 1997] and the GHMMs version of genemark (genemark.hmm, Lukashin and Borodovsky [1998]) and fgenes (fgenesh, Salamov and Solovyev [2000]).

Other gene prediction approaches have been appeared in the same period of time, for instance: mzef [Zhang, 1997], which identified internal coding exons by quadratic discriminant analysis; morgan [Salzberg *et al.*, 1998], which was an integrated system for finding genes in vertebrate DNA sequences by combining different methods with a decision tree classifier; and Augustus [Stanke and Waack, 2003], which incorporated an intron model to an underlying HMM. However, genscan is still considered the standard gene prediction program (at least for human) and it is used in most of the genome annotation pipelines like ENSEMBL and the NCBI genome resources.

3.1.2 Homology based gene-finding

The backbone of similarity-aided or homology-based gene structure determination is constituted by those methods that rely on comparison f the query sequence with protein or cDNA sequences. Database search software, such as BLAST [Altschul *et al.*, 1990, 1997] and related tools, is not capable of automatically identifying start and stop codons or splice sites. Therefore, additional tools are required to define the exonic structures on the potential targets found by the database search programs. Several tools, though, have been developed to calculate spliced alignments, where large gaps—likely to correspond to introns—are only allowed at legal splice junctions, between the query sequence and the to database matches. Among those one can cite SIM4 [Florea *et al.*, 1998], EST_genome [Mott, 1997], Spidey [Wheelan *et al.*, 2001] and exonerate [Slater and Birney, 2005].

Procrustes [Gelfand *et al.*, 1996] and genewise [Birney and Durbin, 1997; Birney *et al.*, 2004b], both predict genes based on a comparison of a genomic query with protein targets. GeneSeqer [Usuka and Brendel, 2000] is a similar spliced alignment program for plant genomes. Projector [Meyer and Durbin, 2004] makes explicit use of the conservation of the exon-intron structure between related genes, which outperforms other tools when the conservation at the amino acid level is weak. Other tools increase the score of candidate exons as a function of the similarity between these exons and known coding sequences resulting of a database search. Examples of this approach are genomescan [Yeh

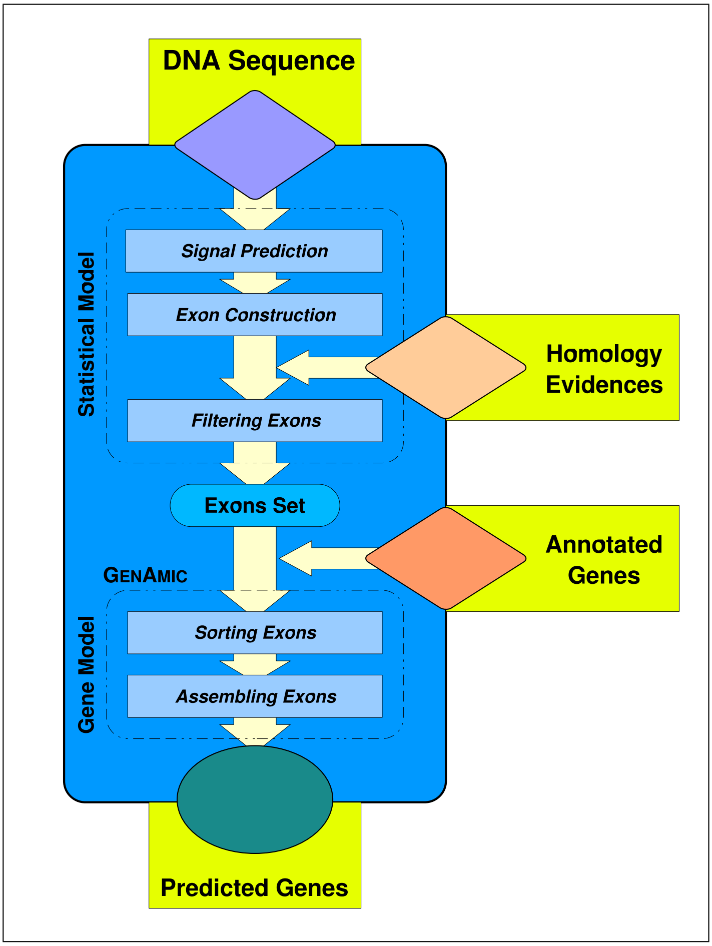


Figure 3.1: **Overall flowchart of geneid.** DNA sequences are scanned to find signals which are then used to build exons. Homology evidences can modify the weights of exons in conserved regions before such exons get filtered to retrieve the high scoring ones. This feature is extensively exploited on SGP2 implementation [Parra *et al.*, 2003]. Those exons are assembled into gene structures by GenAmic, a dynamic programming algorithm with linear asymptotic cost [Guigo, 1998], under a user-defined gene model. At this point, already annotated features can be integrated in the pool of predicted exons. Redrawn from geneid manual figure kindly provided by Enrique Blanco.

et al., 2001], grailexp [Xu *et al.*, 1997] and crasa [Chuang *et al.*, 2003]; the first incorporates similarity to known proteins, the later two use ESTs instead.

3.1.3 Comparative genomics approach

With the availability of many genomes from different species, a number of strategies have been developed to use genome comparisons to predict genes. The rationale behind comparative genomic methods is that functional regions, protein coding regions among them, are more conserved than non-coding ones between genome sequences from different organisms. See, for instance, Figure 3.3 on page 22 (Parra *et al.* 2003, page 109, figure 1) and Figure 5.2 on page 153. This characteristic conservation can be used to identify protein coding exons in the sequences. The approach taken by different programs to exploit this idea differ notably.

In one such approach [Blayo *et al.*, 2002; Pedersen and Scharl, 2002], the problem is stated as a generalization of pairwise sequence alignment: given two genomic sequences coding for homologous genes, the goal is to obtain the predicted exonic structure in each sequence maximizing the score of the alignment of the resulting amino acid sequences. Both Blayo *et al.* [2002] and Pedersen and Scharl [2002] solve the problem through a complex extension of the classical dynamic programming algorithm for sequence alignment. Although very appropriate for short sequences, in practice, the time and memory requirements of this algorithm limit its usefulness for very large genomic sequences. Although the approach theoretically guarantees to produce the optimal amino acid sequence alignment, the fact that sequence conservation may also occur in regions other than protein coding, could lead to overprediction of coding regions, in particular when comparing large genomic sequences from homologous genes from closely related species.

To overcome this limitation, the programs `doublescan` [Meyer and Durbin, 2002] and `SLAM` [Alexandersson *et al.*, 2003] rely on more sophisticated models of coding and non-coding DNA and splice signals, in addition to sequence similarity. Since sequence alignment can be solved with Pair Hidden Markov Models [PHMMs, Durbin *et al.*, 1998] and GHMMs have proven to be very useful to model the characteristics of eukaryotic genes [Burge and Karlin, 1997], `SLAM` and `doublescan` are built upon the so-called Generalized Pair HMMs. In these, gene prediction is not the result of the sequence alignment, as in the programs above, but both gene prediction and sequence alignment are obtained simultaneously.

A third class of programs adopts a more heuristic approach, and separates gene prediction from sequence alignment. The programs `rosetta`[Batzoglou *et al.*, 2000], `SGP1` [from Syntenic Gene Prediction, Wiehe *et al.*, 2001], and `cem` [from the Conserved Exon Method, Bafna and Huson, 2000] are representative of this approach. All these programs start by aligning two syntenic regions (specifically human and mouse in `rosetta`, and `cem`; less species specific in `SGP1`), using some alignment tool (the `glass` program, specifically developed in the case of `rosetta`, or generic ones, such as `TBLASTX`, or `sim96` in the case of `cem` and `SGP1` respectively) and then predict gene structures in which the exons are compatible with the alignment. This compatibility often requires conservation of exonic structure of the homologous genes encoded in the anonymous syntenic regions. Although conservation of exonic structure is an almost universal feature of orthologous human/mouse genes [Waterston *et al.*, 2002], it does not necessarily occur when comparing genomic sequences of homologous genes from other species.

The programs described so far rely on the comparison of fully assembled (and when from different organisms, syntenic) genomic regions. This limits their utility when analyzing complete large eukaryotic genomes and in particular when the informant genome is in non-assembled shotgun form. To overcome this limitation, the programs `Twinscan` [Korf *et al.*, 2001] and `SGP2` [Parra *et al.*, 2003] take a still different approach. The approach in these programs is reminiscent of that used in `genomescan` [Yeh *et al.*, 2001] to incorporate similarity to known proteins to modify the `genscan` scoring schema. Essentially, the query sequence from the target genome is compared against a collection of sequences from the informant genome (which can be a single homologous sequence to the query sequence, a whole assembled genome, or a collection of shotgun reads) and the results of the comparison are used to modify the scores of the exons produced by "*ab initio*" gene prediction programs. In `Twinscan`, the genome sequences are compared using `BLASTN` and the results serve to modify the underlying probability of the potential exons predicted by `genscan`. In `SGP2`, the genome sequences are compared using `TBLASTX`, and the results used to modify the scores of the potential scores predicted by `geneid`; see methods section and Figure 3.4 on page 24 (page 110 and Figure 2 on page 111 of Parra *et al.* 2003).

As the number of available genome sequences of species at different evolutionary distances increases, methods to predict genes based on the comparative analysis of multiple genomes (and not only of two species) look promising. For instance, Dewey *et al.* [2004] combine pairwise predictions from `SLAM` in the human, mouse and rat genomes to simultaneously predict genes with conserved exonic structure in all three species. In the so-called Phylogenetic Hidden Markov Models (phylo-HMMs) or Evolutionary Hidden Markov Models (EHMMs), a gene prediction Hidden Markov Model is combined with a set of evolutionary models, based on phylogenetic trees. Phylo-HMMs take into account that the rate (and type) of evolutionary events differ in protein-coding and non-coding regions. Recently, phylo-HMMs have been applied to gene prediction with encouraging results [Pedersen and Hein, 2003; Siepel and Haussler, 2004].

Phylo-HMMs also have been used in the context of phylogenetic shadowing [Boffelli *et al.*, 2003]. Phylogenetic shadowing examines sequences of closely related species and takes into account the phylogenetic relationship of the set of species analyzed. This approach enables the localization of regions of collective variation and complementary regions of conservation, facilitating the identification of coding as well as non-coding functional regions. The likelihood ratio under a fast (versus slow) mutation regime can be computed for each aligned nucleotide site across all the sequences being analyzed. This ratio represents the relative likelihood that any given nucleotide site was subjected to a faster or slower rate of accumulation of variation and is related to functional constraints imposed on each site. Exon containing sequences will display the least amount of cross species variation, in agreement with the constraint imposed by their function. Regions from different parts of the genome, in which functional non-coding sequences appear, may evolve at different rates [Ebersberger *et al.*, 2002], reflected by differences in their absolute likelihoods. Despite that, functional non-coding regions can be retrieved from stretches of sequence having minimal variation similar to exonic ones.

3.1.4 Analysis pipelines to automatize sequence annotation

Gene prediction software is often integrated into analysis pipelines in order to produce annotations on sets of genomic sequences, for instance a set of chromosome assemblies for a given species or even a bunch of shotgun sequence reads. Here we will shift the focus towards the management of data on which the programs are run and the flow of annotation outputs among different tools. Systems developed to summarize and visualize annotations, that can be incorporated as another step of the annotation process, are extensively described in Chapter 5.

Human annotators use their intuition and experience to synthesize the often contradictory evidence into a single gene structure. Pipelines generally use rules based on the intuition and experience of their designers [Brent and Guigó, 2004]. Human interpretation of the results of these raw analysis by manual curators gives the highest-quality data and most accurate gene structures. However, this process is slow by nature, and annotators may produce conflicting interpretations of the analysis. Fully automated prediction of gene structures has the advantage of being fast, does not require a team of trained annotators, and will process the raw analysis results consistently. Its major drawback, though, is that it can underpredict both the number of genes and the number of alternative transcripts [Potter *et al.*, 2004].

Pise [Letondal, 2001], a web interface generator for molecular biology software, can combine related programs in order to perform more complex analyses. The macros generated by Pise constitute a procedure that will redo the same processing as that already performed, with another initial input. SEALS [Walker and Koonin, 1997] provides a suite of programs designed to facilitate analysis projects involving large amounts of data. The system is designed to provide modular elements which can be combined, modified and integrated with other methods. Pise can be understood as a web interface to analysis programs, while SEALS can be seen as a Unix command-line tool set. However, the first is not meant for automated large-scale analysis, and the latter requires too much manual interaction to be considered a true analysis pipeline.

The ENSEMBL gene-building system [Curwen *et al.*, 2004] enables fast automated annotation of eukaryotic genomes. It annotates genes based on evidence derived from known protein, cDNA and EST sequences. The initial stage of computation is known as the 'raw compute' and comprises various stand-alone analyses, including homology searches using BLAST [Altschul *et al.*, 1997]. Then, ENSEMBL takes these types of analyses one step further and provides a set of gene annotations based on them, to which extra biological information such as gene family, expression data and gene ontologies are linked. Similar systems have been developed for other databases: FLYBASE uses BOP [Mungall *et al.*, 2002], NCBI has its own pipeline [Kitts, 2002] as does the UCSC group [Kent *et al.*, 2002]. The ENSEMBL analysis pipeline [Potter *et al.*, 2004] is split into two parts. The first deals solely with the running of the individual analyses and parsing the output. The second part deals with the automated running, in the correct order, of the many analyses that constitute the pipeline. It keeps track of those that have run succesfully, while also coping with problems such as job failures. In order to scale up the process for the analysis of whole genomes, the pipeline only uses flat files locally on the execution nodes; input data are retrieved directly from a database, and the output data are written back the same way.

Large software systems usually consist of many independently developed parts, and there is a need for data exchange mechanisms to move information among the components. Data integration is a related problem, but with the focus on combining information in scientifically valid ways. Workflow management is the software technology used for keeping track of tasks to be done in generating large datasets or in the automated analysis of such datasets [Goodman, 2002]. A classification of tasks in Bioinformatics emphasizes that most bioinformatics requirements may be described in terms of filters, transformers, transformer-filters, forks and collections of data [Stevens *et al.*, 2001]. Two themes are consistent in these requirements: the need for running analyses in a serial rule-dependent fashion (workflow) and the ability to run these tasks in parallel where possible (highthroughput).

Biopipe [Hoon *et al.*, 2003] is a generic system for large-scale bioinformatics analysis, that has been influenced by the ENSEMBL pipeline. Smaller pipeline systems also exist for annotation of ESTs or individual clones. These systems include Genescript [Hudek *et al.*, 2003] and ASAP [Glasner *et al.*, 2003]. PLAN [Chagoyen *et al.*, 2004] is a simple XML-based language for the definition of executable workflows that simplifies data search and analysis by providing a uniform XML view on both data sources and analytical applications.

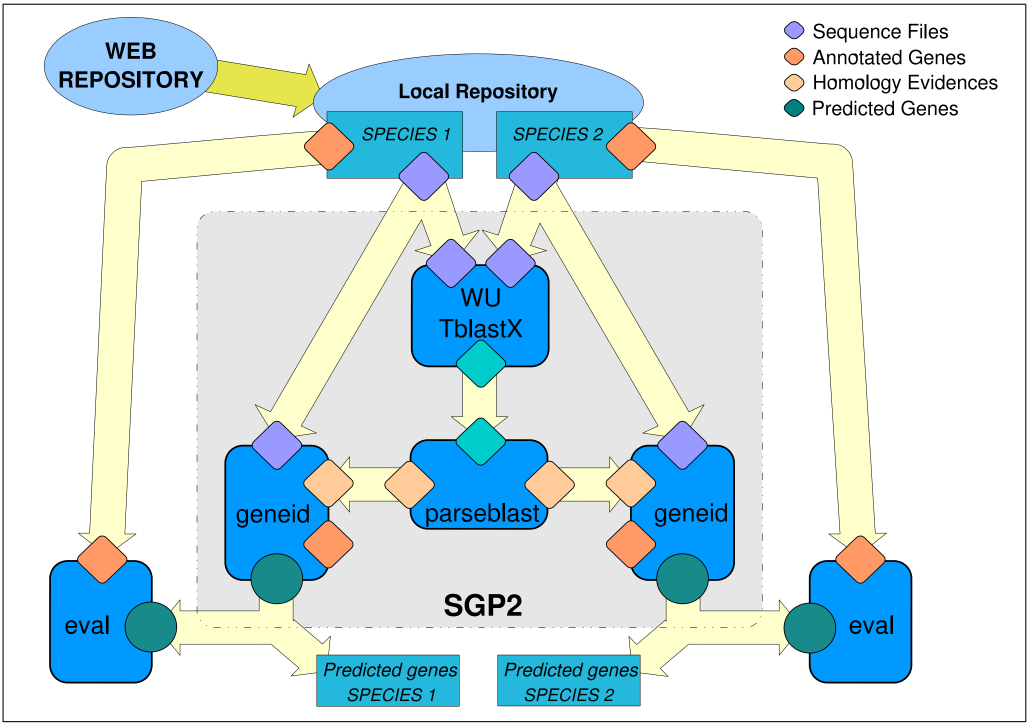


Figure 3.2: **SGP2-based analysis pipeline for pair-wise genome comparisons.** Data is retrieved from a remote server and it is reformatted in the local repository to suit the input for the programs involved in the pipeline. Annotations of known features can be used to train program parameters and to evaluate the outputs for the whole process. In such a scenario, visualizing tools, like gff2ps and gff2aplot (see sections 5.2 and 5.3.1, respectively), can be integrated in the pipeline to summarize predicted genes and homology features.

3.2 SGP2: Syntenic Gene Prediction Tool

The computational approach to incorporate information from the comparison of two genomes to geneid is described in the research article attached in the following subsection (see Section 3.2.1 on page 20), and it was briefly discussed on page 15 of Section 3.1.3. Here, we would like to discuss SGP2 in the context of the genomic comparison between human and mouse, which is reflected in Section 3.2.2, page 31. The results for those analyses are summarized in Section "*De novo* gene prediction" on page 38 (page 539 of Waterston *et al.* 2002).

Figure 3.2 on page 18 displays a general analysis protocol to produce a set of gene predictions in a set of sequences for different species. SGP2 can be seen there as a procedure based on TBLASTX and geneid. It also requires some programs to filter the similarity regions found by TBLASTX. In the figure, only the parseblast filter was drawn for the sake of simplicity, but there are a few other programs involved in the SGP2 processing of similarity data. The algorithms and the parameter settings for the software are detailed in methods section and Figure 3.4 on page 24 (page 110 and Figure 2 on page 111 of Parra *et al.*

2003)

A whole analysis pipeline was developed for the human and mouse genome comparisons. It included preprocessing of the genome sequences and annotations from the UCSC FTP server; the search for homology between the sequences of the two genomes; the computational gene prediction approaches, both the "*ab initio*" (`geneid` and `genscan`) and the comparative genomics approach (SGP2). Results from other groups were also integrated in the pipeline in order to perform the evaluation of the gene predictions against different reference annotation sets (including REFSEQ and ENSEMBL genes). At that time there were updates of sequence sets for each genome version that was assembled for the human and mouse genomes. This required to run again the whole protocol on those new genomic sequences. Another issue was the growing number of elements to be included in the analysis pipeline. To face both problems, we developed a simple task manager in `perl` to control the processes to be run on a given set of sequences, and to distribute the task among different machines of our lab. The `perl` program was provided with a set of unix shell scripts to be run in a given order and with a set of sequences. It scheduled all the jobs to be run for each sequence by using a simple execution queue. The task manager sent each job script to be executed on a sequence to a machine in the list of available computers of our lab. This was achieved with `rsh` remote shell calls, while the sequence files and the results were shared among all the computers involved in the analysis via the Network File System (NFS). The task scheduler also kept a record of the execution status of each submitted job, reporting those cases in which the remote execution failed, without resubmitting them.

The major drawback of this simple approach was the bottleneck of using flat files throught the NFS on multiple computers when programs required intensive input/output flow to the file system. This has been already stated in Potter *et al.* [2004], and was the reason for the development of the ENSEMBL analysis pipeline with a relational database system. However, the modular design of the shell scripts defining each job warranted that many of the components of the semi-automated analysis pipeline described in this section were recycled. They have been used to obtain predictions for new versions of the human and mouse genomes, but also for other genomes of species such as rat and chicken. The results have been collected in a web repository (see the "Gene Predictions on Genomes" entry in the *Web Glossary*, on page 214).

3.2.1 Parra *et al*, *Genome Research*, 13(1):108–117, 2003

PubMed Accession:

http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&list_uids=12529313&dopt=Abstract

Journal Abstract:

http://www.genome.org/cgi/content/abstract/13/1/108

Supplementary Materials:

http://genome.imim.es/datasets/sgp2002/

Program Home Page:

http://genome.imim.es/software/sgp2/

Parra G, Agarwal P, Abril JF, Wiehe T, Fickett JW, Guigo R.

Comparative gene prediction in human and mouseGenome Research.

2003 Jan;13(1):108-17.

http://www.genome.org/cgi/content/full/13/1/108

Methods

Comparative Gene Prediction in Human and Mouse

Genís Parra,[1] Pankaj Agarwal,[2] Josep F. Abril,[1] Thomas Wiehe,[3] James W. Fickett,[4] and Roderic Guigó[1,5]

[1]Grup de Recerca en Informàtica Biomèdica. Institut Municipal d'Investigació Medica / Universitat Pompeu Fabra / Centre de Regulació Genòmica 08003 Barcelona, Catalonia, Spain; [2]GlaxoSmithKline, King of Prussia, Pennsylvania 19406, USA; [3]Freie Universität Berlin and Berlin Center for Genome Based Bioinformatics (BCB), 14195 Berlin, Germany; [4]AstraZeneca R&D Boston, Waltham, Massachusetts 02451, USA

The completion of the sequencing of the mouse genome promises to help predict human genes with greater accuracy. While current ab initio gene prediction programs are remarkably sensitive (i.e., they predict at least a fragment of most genes), their specificity is often low, predicting a large number of false-positive genes in the human genome. Sequence conservation at the protein level with the mouse genome can help eliminate some of those false positives. Here we describe SGP2, a gene prediction program that combines ab initio gene prediction with TBLASTX searches between two genome sequences to provide both sensitive and specific gene predictions. The accuracy of SGP2 when used to predict genes by comparing the human and mouse genomes is assessed on a number of data sets, including single-gene data sets, the highly curated human chromosome 22 predictions, and entire genome predictions from ENSEMBL. Results indicate that SGP2 outperforms purely ab initio gene prediction methods. Results also indicate that SGP2 works about as well with 3x shotgun data as it does with fully assembled genomes. SGP2 provides a high enough specificity that its predictions can be experimentally verified at a reasonable cost. SGP2 was used to generate a complete set of gene predictions on both the human and mouse by comparing the genomes of these two species. Our results suggest that another few thousand human and mouse genes currently not in ENSEMBL are worth verifying experimentally.

After the genome sequence of an organism has been obtained, the very first next step is to compile a complete and accurate catalog of the genes encoded in this sequence. For higher eukaryotic organisms, however, the accuracy of currently available gene prediction methods to perform such a task is limited (Guigó et al. 2000; Rogic et al. 2001; Guigó and Wiehe 2003). The increasing availability of genome sequences from different organisms, however, has lead to the development of new computational gene finding methods that use sequence conservation to help identifying coding exons, and improve the accuracy of the predictions (Fig. 1; Crollius et al. 2000; Wiehe et al. 2000; Miller 2001; Rinner and Morgenstern 2002). Indeed, three such comparative gene prediction programs, SLAM (Pachter et al. 2002), SGP2, and TWINSCAN (Korf et al. 2001) have been used for the comparative analysis of the human and mouse genomes. These analyses lead to more accurate gene predictions, and to the verification of previously unconfirmed genes. In this paper, we describe the program SGP2. Typical computational ab initio gene prediction methods rely on the identification of suitable splicing sites, start and stop codons along the query sequence, and the computation of some measure of coding likelihood to predict and score candidate exons, and delineate gene structures (see Claverie 1997; Burge and Karlin 1998; Haussler 1998; Zhang 2002 and references therein for reviews on computational gene finding).

Similarity between the query sequence and known coding sequences (amino acid or cDNA) can also be used to infer gene structures. When the query sequence encodes a protein for which a close homolog exists, a special type of alignment can be used between the DNA sequence and the target protein/cDNA sequence, in which gaps in the target sequence corresponding to introns in the query sequence must be compatible with potential splicing signals. This is the approach in GENEWISE (Birney and Durbin 1997) and PROCRUSTES (Gelfand et al. 1996). Alternatively, the results of searching the query sequence against a database of known coding sequences, using for instance BLASTX (Altschul et al. 1990, 1997; Gish and States 1993), can be incorporated more or less ad hoc into the scoring schema of an ab initio gene prediction method. The program GENOMESCAN (Yeh et al. 2001), which incorporates BLASTX search results into the predictions by the GENSCAN program (Burge and Karlin 1997), is an example of a recent development in that direction.

Recently developed comparative gene prediction programs further exploit sequence similarity. Instead of comparing anonymous genomic sequences to known coding sequences, anonymous genomic sequences are compared to anonymous genomic sequences from the same or different organisms, under the assumption that regions conserved in the sequence will tend to correspond to coding exons from homologous genes. The approach taken by the different programs to exploit this idea differs notably.

In one such approach (Blayo et al. 2002; Pedersen and Scharl 2002), the problem is stated as a generalization of pairwise sequence alignment: Given two genomic sequences coding for homologous genes, the goal is to obtain the predicted exonic structure in each sequence maximizing the score of the

[5]**Corresponding author.**
E-MAIL rguigo@imim.es; FAX 34 93 224-0875.

Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.871403.

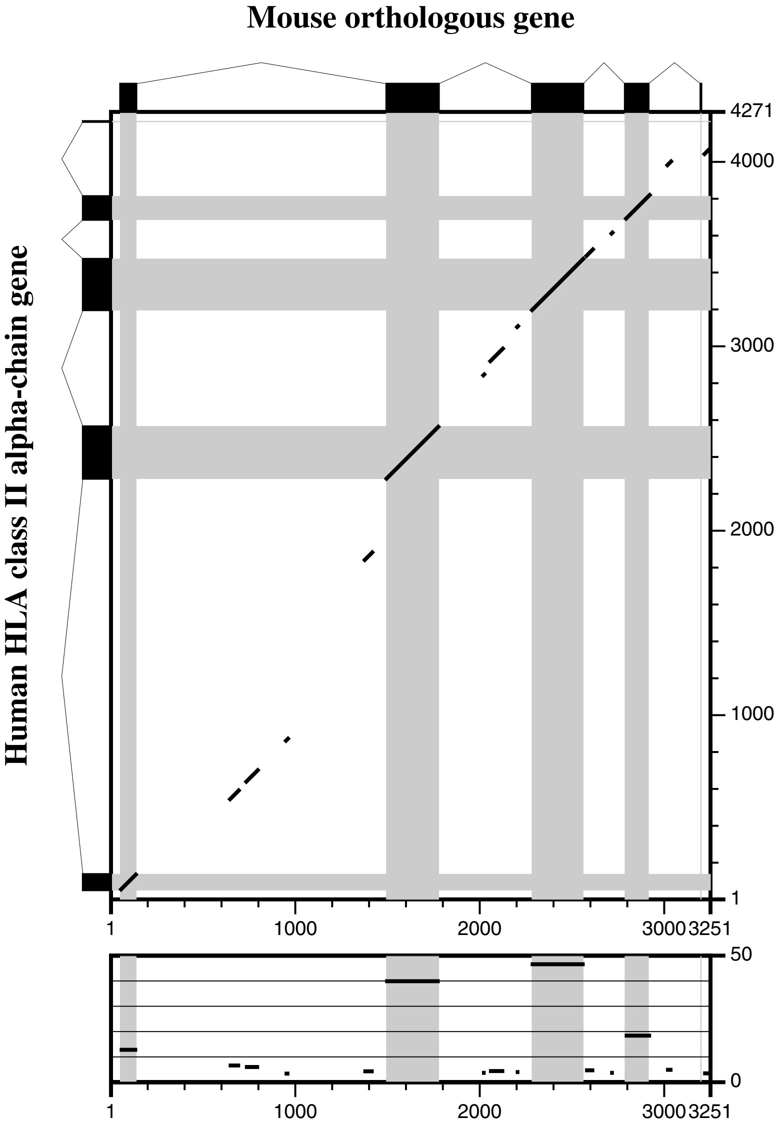


Figure 1 Pairwise comparison using TBLASTX of the human and mouse genomic sequences coding for the HLA class II alpha chain. Black boxes indicate the coding exons, while black diagonals indicate the conserved alignments. The score of the conserved alignments (divided by 10) is given in the lower panels. Although conserved regions between the human and mouse genomic sequences coding for these genes fully include the coding exons, a substantial fraction of intronic regions is also conserved. The TBLASTX outptut was post-processed to show a continuous non-overlapping alignment.

alignment of the resulting amino acid sequences. Both Blayo et al. (2002) and Pedersen and Scharl (2002) solve the problem through a complex extension of the classical dynamic programming algorithm for sequence alignment.

In a different approach, the programs SLAM (Pachter et al. 2002) and DOUBLESCAN (Meyer and Durbin 2002) combine sequence alignment pair hidden Markov Models (HMMs; Durbin et al. 1998) with gene prediction generalized HMMs (GHMMs; Burge and Karlin 1997) into the so-called generalized pair HMMs. In these, gene prediction is not the result of the sequence alignment, as in the programs above; gene prediction and sequence alignment are obtained simultaneously.

A third class of programs adopt a more heuristic approach, and separate clearly gene prediction from sequence alignment. The programs ROSSETA (Batzoglou et al. 2000), SGP1 (from 'syntenic gene prediction'; Wiehe et al. 2001), and CEM (from 'conserved exon method'; Bafna and Huson 2000) are representative of this approach. All these programs start by aligning two syntenic sequences and then predict gene structures in which the exons are compatible with the alignment. The programs described thus far rely on the comparison of fully assembled (and when from different organisms, syntenic) genomic regions. This limits their utility when analyzing complete large eukaryotic genomes, and in particular when the informant genome is in nonassembled shotgun form. To overcome this limitation, the programs TWINSCAN (Korf et al. 2001) and SGP2 take still a different approach. The approach is reminiscent of that used in GENOMESCAN (Yeh et al. 2001) to incorporate similarity to known proteins to modify the GENSCAN scoring schema. Essentially, the query sequence from the target genome is compared against a collection of sequences from the informant genome (which can be a single homologous sequence to the query sequence, a whole assembled genome, or a collection of shotgun reads), and the results of the comparison are used to modify the scores of the exons produced by ab initio gene prediction programs. In TWINSCAN, the genome sequences are compared using BLASTN, and the results serve to modify the underlying probability of the potential exons predicted by GENSCAN. In SGP2, the genome sequences are compared using TBLASTX (W. Gish, 1996–2002, http://blast.wustl.edu), and the results are used to modify the scores of the potential scores predicted by GENEID. TWINSCAN and SGP2 have been successfully applied to the annotation of the mouse genome

(Mouse Genome Sequencing Consortium 2002), and have helped to identify previously unconfirmed genes (Guigó et al. 2003).

In the next section, we describe the algorithmic details of SGP2, and its implementation. We also describe the sequence sets used to benchmark SGP2 accuracy. Results based on these data sets indicate that SGP2 is an improvement over pure ab initio gene prediction programs, even when the informant genome is only in shotgun form. We have found that 3x coverage will generally suffice to achieve maximum accuracy. Finally, we describe the application of SGP2 to the comparative analysis of the human and mouse genomes.

METHODS

SGP2

SGP2 is a method to predict genes in a *target* genome sequence using the sequence of a second *informant* or *reference* genome. Essentially, SGP2 is a framework to integrate the ab initio gene prediction program GENEID (Guigó et al. 1992; Parra et al. 2000) with the sequence similarity search program TBLASTX. The approach is conceptually similar to that used in TWINSCAN to incorporate BLASTN searches into GENSCAN.

GENEID is a genefinder that predicts and scores all potential coding exons along a query sequence. Scores of exons are computed as log-likelihood ratios, which are a function of the splice sites defining the exon, and of the coding bias in composition of the exon sequence as measured by a Markov Model of order five (Borodovsky and McIninch 1993). From the set of predicted exons, GENEID assembles the gene structure (eventually multiple genes in both strands), maximizing the sum of the scores of the assembled exons, using a dynamic programming chaining algorithm (Guigó 1998).

When using an informant genome sequence to predict genes in a target genome sequence, ideally we would like to incorporate into the scores of the candidate exons predicted along the target sequence, the score of the optimal alignment at the amino acid level between the target exon sequence and the counterpart homologous exon in the informant genome sequence. If a substitution matrix, for instance from the BLOSUM family, is used to score the alignment, the resulting score can also be assumed to be a log-likelihood ratio: informally, the ratio between the likelihood of the alignment when the amino acid sequences code for functionally related proteins, and the likelihood of the alignment, otherwise. In principle, this score could be added to the GENEID score for the exon. TBLASTX provides an appropriate shortcut to often find a good enough approximation to such an optimal alignment, and infer the corresponding score: The optimal alignment can be assumed to correspond to the maximal scoring high-scoring segment pairs (HSP) overlapping the exon. However, when dealing in particular with the informant genome sequence in fragmentary shotgun form, often different regions of a candidate exon sequence will align optimally to different informant genome sequences. Thus, in the approach used here, we identify the optimal HSPs covering each fraction of the exon, and compute separately the contribution of each HSP into the score of the exon. In the next section, we describe in detail how this computation is performed.

Scoring of Candidate Exons

Let e be one of the candidate exons predicted by GENEID along the query DNA sequence S. In SGP2, the final score of e, $s(e)$, is computed as

$$s(e) = s_g(e) + ws_t(e)$$

where $s_g(e)$ is the score given by GENEID to the exon e, and $s_t(e)$ is the score derived from the HSPs found by a TBLASTX search overlapping the exon e. Both scores are log-likelihood ratios (and we compute both base two). Assuming that both components are independent, they can be summed up into a single score. However, the assumption of independence is not realistic, $s_g(e)$ depends on the probability of the sequence of e, assuming that e codes for a protein, while $s_t(e)$ depends on the probability of the optimal alignment of e with a sequence fragment of the mouse genome, assuming that both sequences code for related proteins. Obviously, these two probabilities are not independent. Their joint distribution could only be investigated—at least empirically—if the Markov Model of coding DNA used in GENEID, and the substitution matrix used by TBLASTX were inferred from the very same set of coding sequences. Since this is quite difficult, if not unfeasible, we use an "ad hoc" coefficient, w, to weight the contribution of TBLASTX search, $s_t(e)$ into the final exon score.

We compute $s_t(e)$ in the following way. Let $h_1 \cdots h_q$ be the set of HSPs found by TBLASTX after comparing the query sequence S against a database of DNA sequences (Fig. 2A).

First, we find the *maximum scoring projection* of the HSPs onto the query sequence. We simply register the maximum score among the scores of all HSPs covering each position, and then partition the query sequence in equally maximally scoring segments (bounded by dotted lines in Fig. 2A) $x_1 \cdots x_r$, with scores $s_p(x_1) \cdots s_p(x_r)$ (Fig. 2B).

Then, for each predicted exon e (Fig. 2C), we find X_e, the set of maximally scoring segments overlapping e

$$X_e = \{x_i : x_i \cap e \neq \emptyset\}$$

where $a \cap b$ denotes the overlap between sequence segments a and b, and $\emptyset$ means no overlap. We compute $s_t(e)$in the following way:

$$s_t(e) = \sum_{x \in X_e} s_p(x) \frac{|x \cap e|}{|x|}$$

where $|a|$ denotes the length of sequence segment a.

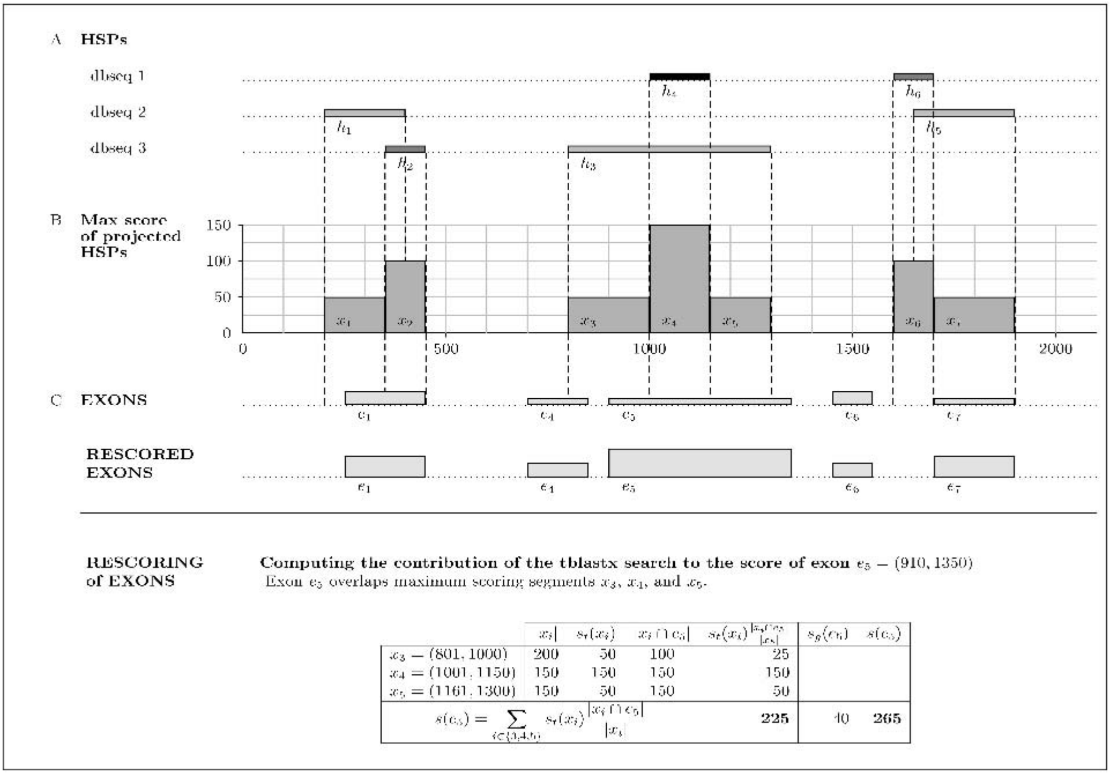
That is, each exon gets the score of the maximally scoring HSPs along the exon sequence proportional to the fraction of the HSP covering the exon. In other words, $s_t(e)$ is the integral of the maximum scoring projection function within the exon interval.

Once the scores s have been computed for all predicted exons in the sequence S, gene prediction proceeds as usual in GENEID: The gene structure is assembled maximizing the sum of scores of the assembled exons.

Running SGP2

In practice, we run SGP2 in the following way. Given a DNA query sequence and a collection of DNA sequences, we compare the query sequence against the collection using TBLASTX 2.0MP-WashU [23-Sep-2001]. The query sequence can be a genomic fragment of any size, including complete eukaryotic chromosomes, whereas the collection of sequences may be almost anything from just a homologous region or a partial collection of genomic sequences from the same or another species to the whole genome sequence of a second species, either completely assembled or in shotgun form at any degree of coverage. In particular, two different regions of the same genome coding for homologous genes can be used within SGP2; in this case the same genome acts as target and informant.

In all the analyses reported here, we used BLOSUM62 as the amino acid substitution matrix, but changed the penalty for aligning any residue to a stop codon to -500. This helps to get rid of a large fraction of HSPs in noncoding regions. Because of TBLASTX limitations, large query sequences may need to be split in fragments before the search, and the results reconstructed afterwards. Results of TBLASTX search are then



Computing the contribution of the tblastx search to the score of exon e_5 – (910, 1350)
Exon e_5 overlaps maximum scoring segments x_3, x_4, and x_5.

	$\|x_i\|$	$s_t(x_i)$	$\|x_i \cap e_5\|$	$s_t(x_i)\frac{\|x_i \cap e_5\|}{\|x_i\|}$	$s_g(e_5)$	$s(e_5)$
x_3 – (801, 1000)	200	50	100	25		
x_4 = (1001, 1150)	150	150	150	150		
x_5 = (1161, 1300)	150	50	150	50		
$s(e_5) = \sum_{i \in \{3,4,5\}} s_t(x_i)\frac{\|x_i \cap e_5\|}{\|x_i\|}$				**225**	40	**265**

Figure 2 Rescoring of the exons predicted by GENEID according to the results of a TBLASTX search. See the "SGP2" section for a detailed explanation of the figure.

parsed to obtain the *maximum scoring projection* of the HSPs onto the query sequence. The parsing includes discarding all HSPs below a given bit score cutoff, subtracting this value from the score of the remaining HSPs, weighting the resulting score by *w* (see above), and collapsing the HSPs in to the maximum scoring projections. In all analyses described here, the bit score cutoff was set to 50, and *w* to 0.20. These values were chosen to optimize the gene predictions in sequence sets of known homologous human and mouse genomic sequences (see the Results section).

The *maximum scoring projection* is given to GENEID in general feature format (GFF; R. Durbin and D. Haussler, http://www.sanger.ac.uk/Software/GFF/). GENEID uses it to rescore the exons predicted along the query sequence as explained, and assembles the corresponding optimal gene structure. GENEID was already designed to incorporate external information into the gene predictions, and no changes were required in the program to accommodate it into the SGP2 context, only a small adjustment in the parameter file to cope with the change in scale of the exon scores.

We have written a simple PERL script which, given a query DNA sequence and the results of the TBLASTX search, performs all the components of the SGP2 analysis transparently: the parsing of the TBLASTX search results, and the GENEID predictions. In the case wherein both the query and the informant sequence are single genomic fragments, the gene predictions can be obtained in both sequences (without the need for a second TBLASTX search). The script, as well as the individual components, can be found at http://www1.imim.es/software/sgp2/.

GENEID has essentially no limits to the length of the input sequence, and deals well with chromosome size sequences. Limits to the length of the input query sequence that can be analyzed by SGP2 are, thus, those imposed by TBLASTX. GENEID is quite fast; given the parsed TBLASTX results, it takes 6 h to reannotate the whole human genome in a MOSIX cluster containing four PCs (PentiumIII Dual 500 Mhz processors).

Accelerating TBLASTX Searches

TBLASTX searches, although efficient, are much slower. Its default usage may become computationally prohibitive when comparing complete eukaryotic genomes. In the context of SGP2, however, a number of TBLASTX options can be changed to speed up the search, without significant loss of sensitivity in the predictions (see the Results section). Thus, results in human chromosome 22 and whole-genome comparisons have been performed using the following set of parameters: W = 5, -nogap, -hspmax = 150,000, B = 200, V = 200, E = 0.01, E2 = 0.01, Z = 30,000,000, -filter = xnu + seg, and S2 = 80. In these cases, the query sequences have been broken up in 5 MB fragments, and the database sequences in 10 MB fragments. In all cases, stop codons are heavily penalized (−500) in the alignments. After the search is completed, locations of the resulting HSPs are recomputed in chromosomal coordinates. Results in the single-gene sequence benchmark data sets were obtained with default TBLASTX parameters.

Sequence Data Sets

Benchmark Sequence Sets

To optimize some of the parameters in SGP2 and to test its performance, we used a set of known pairs of genomic sequences coding for homologous human and rodent genes. The set is built after the set constructed by Jareborg et al. (1999). This is a set of 77 orthologous mouse and human gene pairs. We considered only the 33 pairs of sequences in this set coding for single complete genes. In addition, we discarded six additional pairs, when we suspected that one of the members could be wrongly annotated. Orthology in the Jareborg et al. (1999) data set is based on sequence conservation. This could bias the set towards the more highly conserved human/mouse orthologous genes. To compensate for this bias, we obtained an additional set of pairs of human/rodent orthologous genes through an approach which does not involve sequence conservation: We obtained the set of pairs of human/mouse sequences from the SWISSPROT database sharing the prefix (indicating the gene) in their locus names. We kept only those pairs for which it was possible to find the corresponding annotated genomic sequence—including the mapping of the transcript, and not only of the coding regions—in the EMBL database. Fifteen additional genes were found this way. Three of them were discarded because we suspected wrong annotation in at least one of the members of the pair. We believe that orthology in the remaining cases is highly likely because of the absolute conservation of the exonic structure (number and length of exons, and intron phases) that we observed. We will call the resulting concatenated set of 39 pairs of human/mouse homologous genes the SCIMOG dataset (from Sanger Center IMim Orthologous Genes). The data set and the detailed protocol used to obtain it can be accessed at http://www1.imim.es/datasets/sgp2002/.

To test the accuracy of SGP2, we used the data set constructed by Batzoglou et al. (2000) of 117 orthologous human and mouse genes. We discarded those pairs in which in at least one of the sequences contained multiple genes, and those in which the coding region started in position 1 in one of the sequences of the pair. This resulted in 110 genes. We will call this set the MIT data set. There is some overlap between the SCIMOG and MIT data sets, and thus the latter cannot properly be called a test set. However, we decided not to eliminate the redundant entries, so that the results could be compared to those published for the ROSSETA program (Batzoglou et al. 2000).

Finally, we tested SGP2 in the complete sequence of human chromosome 22 (Dunham et al. 1999). The masked sequence was obtained from http://genome.cse.ucsc.edu/goldenPath/22dec2001/. Chromosome 22 is probably the best annotated human chromosome. We used the gene annotations at http://www.cs.columbia.edu/~vic/sanger2gbd/. The CDS set contains 554 genes. This is a conservative set that only contains the coding region of genes and does not include pseudogenes. This may lead to an underestimation of the specificity of the predictions.

Mouse and Human Genome Sequences

We used versions MGSCv3 of the mouse genome (2,726,995,854 bp, http://genome.cse.ucsc.edu/goldenPath/mmFeb2002/) and NCBI28 of the human genome (3,220,912,202 bp, http://genome.cse.ucsc.edu/goldenPath/22dec2001/). Both masked and unmasked sequences were obtained from these locations. ENSEMBL gene annotations for these genomes were obtained from http://genome.cse.ucsc.edu/goldenPath/22dec2001/database/ensGene.txt.gz for the human genome, and from http://genome.cse.ucsc.edu/goldenPath/mmFeb2002/database/ensGene.txt.gz for the mouse genome. ENSEMBL predicts 23,005 and 22,076 nonoverlapping transcripts genes on the human and mouse genome, respectively.

Evaluating Accuracy

The measures of accuracy used here are extensively discussed in Burset and Guigó (1996). We will restate them briefly. Accuracy is measured at three different levels: nucleotide, exon, and gene. At the nucleotide and exon levels, we compute essentially the proportion of actual coding nucleotides/exons that have been correctly predicted—which we call *sensitivity*—and the proportion of predicted coding nucleotides/exons that are actually coding nucleotides/exons—which we call *specificity*. To compute these measures at the exon level, we will assume that an exon has been correctly predicted only when both its boundaries have been correctly predicted. To summarize both *sensitivity* and *specificity*, we compute the *correlation coefficient* at the nucleotide level, and the average of *sensitivity* and *specificity* at the exon level. At the exon level, we also compute the *missing exons*, the proportion of actual exons that overlap no predicted exon, and the *wrong exons*, the proportion of predicted exons that overlap no real exons.

At the gene level, a gene is correctly predicted if all of the coding exons are identified, every intron–exon boundary is correct, and all of the exons are included in the proper gene. In addition, we compute the missed genes (MGs), real genes for which none of its exons are overlapped by a predicted gene, and the wrong genes (WGs), predictions for which none of the exons are overlapped by a real gene. In general, gene finders predict the initial and terminal exons very poorly. This often leads to so-called chimeric predictions—one predicted gene encompassing more than one real gene—or to split predictions—one real gene split in multiple predicted genes. Reese et al. (2000) developed two measures, split genes (SG) and joined genes (JG), to account for these tendencies. SG is the total number of predicted genes overlapping real genes divided by the number of genes that were split. Similarly, JG is the total number of real genes that overlap predicted genes divided by the number of predicted genes that were joined.

RESULTS

Benchmarking SGP2

We evaluated the accuracy of SGP2 using a number of different data sets. The lack of a gold standard of gene prediction makes it difficult to get accurate assessments from any single data set. We primarily used three data sets as described earlier.

To benchmark SGP2, we constructed BLAST databases from the mouse and human sections of SCIMOG and MIT, and each mouse/human sequence to the entire human/mouse database, respectively. This enabled us to predict genes in both the mouse and human databases. The results from

Table 1. Gene Prediction in the SCIMOG Data Set

	Nucleotide			Exon				
Program	Sn	Sp	CC	Sn	Sp	(Sn+Sp)/2	ME	WE
GENSCAN	0.98	0.86	0.92	0.84	0.75	0.79	0.04	0.14
TBLASTX default	0.89	0.76	0.81	0.81	—	—	0.19	0.11
SGP2 (single complete genes)	0.97	0.98	0.97	0.89	0.89	0.89	0.03	0.03
SGP2 (multiple genes)	0.94	0.97	0.95	0.80	0.87	0.83	0.10	0.02

Table 2. Gene Prediction Accuracy in the MIT Data Set

	Nucleotide			Exon				
Program	Sn	Sp	CC	Sn	Sp	(Sn+Sp)/2	ME	WE
GENSCAN	0.98	0.89	0.93	0.82	0.75	0.78	0.06	0.13
ROSSETA	0.95	0.97	—	—	—	—	0.02	0.03
TBLASTX default	0.94	0.79	0.85	—	—	—	0.13	0.13
SGP2 (single complete genes)	0.97	0.98	0.97	0.84	0.85	0.84	0.05	0.03
SGP2 (multiple genes)	0.96	0.97	0.96	0.71	0.79	0.75	0.12	0.03

comparing SGP2, GENSCAN, and ROSSETA accuracy values in this case are taken from Batzoglou et al. (2000), and the results of a simple TBLASTX search on the MIT data set are in Table 2 (below). For the TBLASTX searches, the *maximum scoring projection* of the HSPs (see the above section titled "SGP2") was assumed to be the gene prediction. The score cutoff for the HSPs was chosen to maximize the correlation coefficient (CC) between the projected HSPs and the coding exons. In Table 1,2, we report the accuracy of GENSCAN, SGP2, and TBLASTX on the SCIMOG dataset. The accuracy values for SGP2 are reported under two scenarios: assuming a single complete gene and assuming multiple genes. Both GENEID and SGP2 allow the external specification of a *gene model* (i.e., a small number of rules specifying the legal assemblies of exons into gene structures). These rules can be used to force SGP2 to predict a single complete gene to make the results comparable to those of ROSSETA. Without such a restriction (i.e., making no assumptions about the number and completeness of the genes potentially encoded in the query sequence), the results are more directly comparable to those of GENSCAN (although GENSCAN also has a tendency to start a prediction in any sequence with an initial exon, and to terminate it with a terminal exon).

The accuracy of SGP2 is comparable to that of ROSSETA, and is significantly higher than that of GENSCAN. SGP2 also improves substantially over a simple TBLASTX search. The relative low specificity of the TBLASTX search—even after the large penalties for stop codons—reflects the fact that a substantial fraction of the conservation between the human and mouse genomes extends into the noncoding regions (Mouse Genome Sequencing Consortium 2002). At the nucleotide level, SGP2 accuracy is almost equal in the MIT data set and the SCIMOG data set (even though the SGP2 was trained on SCIMOG). The accuracy at the exact exon level, however, decreases, in particular when prediction of multiple genes is allowed. This is a problem inherited from GENEID, which tends to replace short initial and terminal exons with longer internal exons.

Accuracy of SGP2 as a Function of the Coverage of the Mouse Genome

To investigate the utility of partial shotgun data as informant sequence in our approach based on TBLASTX, we simulated shotgun mouse sequence data at different levels of coverage (1.5x, 3x, and 6x) from the mouse genes in the SCIMOG data set, and used them to compare the human sequences in SCIMOG using TBLASTX. The mouse genomic sequences was shredded with uniformly distributed length between 500 and 600 bp with random starting points. No sequencing errors were introduced. At each coverage, we measured the CC between the TBLASTX hits projected along the human genome sequence, and the coding exons (choosing the TBLASTX score cutoff resulting in the optimal CC). With 1.5x coverage, a substantial fraction of the human coding region is not identified by TBLASTX, whereas with 3x, the results are quite similar to those obtained with 6x, which are identical to those obtained with the fully assembled syntenic regions (Table 3). This indicates that even with 3x coverage of the informant genome, our method will produce results nearly identical to those obtained with fully assembled regions. Assembled genomes, however, result in faster TBLASTX searches.

Accuracy of SGP2 in Human Chromosome 22

Human chromosome 22 was the first human chromosome fully sequenced (Dunham et al. 1999), and it is quite the best annotated thus far, due to a number of experimental followups (Das et al. 2001; Shoemaker et al. 2001). Therefore, it provides an excellent data set to validate any gene prediction technology. Human chromosome 22 was searched using TBLASTX against the masked whole-genome assembly from the mouse genome (MGSCv3). The HSPs in chromosomal coordinates resulting from the TBLASTX search were used in GENEID to perform SGP2 gene prediction. Although the HSPs had been computed on the masked sequence, in this case the SGP2 predictions were obtained on the unmasked one. SGP2 predicted 729 genes on human chromosome 22. Table 4 shows the comparative accuracy of the SGP2, GENSCAN, GENOMESCAN, and pure ab initio GENEID predictions (without TBLASTX data). GENSCAN predictions on the masked sequence were taken from the USCS genome browser http://genome.cse.ucsc.edu/. GENOMESCAN predictions were obtained from ftp://ftp.ncbi.nih.gov/genomes/H_sapiens/build28_chr_genomescan.gtf.gz. Pure ab initio GENEID predictions were obtained on the masked sequence, and can also be downloaded from http://www1.imim.es/genepredictions/.

Although SGP2 is not more sensitive than GENSCAN, it appears to be more specific (as it utilizes the mouse genome).

Table 3. Accuracy of TBLASTX Predictions as a Function of the Degree of Coverage in the SCIMOG Data Set

	Nucleotide			Exon	
Coverage	Sn	Sp	CC	ME	WE
Simulated 1.5x	0.79	0.78	0.77	0.25	0.10
Simulated 3x	0.86	0.76	0.80	0.21	0.11
Simulated 6x	0.89	0.76	0.81	0.19	0.11
Fully assembled	0.89	0.76	0.81	0.19	0.11

Table 4. Accuracy of Gene-finding Programs on Human Chromosome 22

	Nucleotide			Exon					Gene						
Program	Sn	Sp	CC	Sn	Sp	(Sn+Sp)/2	ME	WE	Sn	Sp	(Sn+Sp)/2	MG	WG	JG	SG
GENSCAN	0.86	0.50	0.64	0.70	0.40	0.55	0.13	0.50	0.06	0.04	0.05	0.11	0.45	1.24	1.07
GENOMESCAN	0.87	0.44	0.59	0.72	0.36	0.54	0.10	0.55	0.11	0.06	0.08	0.12	0.52	1.07	1.14
GENEID	0.80	0.63	0.69	0.66	0.53	0.59	0.19	0.35	0.09	0.07	0.08	0.14	0.39	1.20	1.08
TBLASTX	0.84	0.39	0.54	—	—	—	0.12	0.74	—	—	—	0.11	—	—	—
SGP2	0.83	0.67	0.73	0.68	0.56	0.62	0.16	0.31	0.13	0.10	0.11	0.14	0.36	1.14	1.13

Fifty percent of the GENSCAN-predicted exons do not overlap annotated chromosome 22 exons; this number is only 31% for SGP2. Overall, SGP2 appears to be more accurate than GENSCAN in human chromosome 22: GENSCAN's CC at the nucleotide level is 0.64, whereas that of SGP2 is 0.73. Although accuracy decreases for both programs when going from single-gene sequences (Tables 1, 2) to an entire chromosome, SGP2 retains more accuracy. GENSCAN overall shows higher sensitivity than SGP2, but there were 45 real genes not predicted by GENSCAN on human chromosome 22, and SGP2 was able to predict, at least partially, 15 of them. This suggests that SGP2 and GENSCAN may play complementary roles. GENOMESCAN, on the other hand, did not appear to be superior to GENSCAN in human chromosome 22.

Mouse matches (TBLASTX HSPs) covered 11% of the human chromosome 22. Though they covered 85% of the coding nucleotides, 74% of the HSPs fell outside annotated coding regions. This illustrates the difficulties of using genome sequence conservation even at the protein level between human and mouse genomes to infer coding genes.

Prediction of Genes in the Human and Mouse Genomes

We used SGP2 to predict the entire complement of human (NCBI28) and mouse (MGSCv3) genes. The masked sequences of these two genomes were compared using TBLASTX. The TBLASTX HSPs were used within SGP2. SGP2 predicted 44,242 genes in the human genome, and 44,777 genes in the mouse genome. Obviously, it is difficult to accurately assess these predictions. We used ENSEMBL genes as the set of reference annotations and compared both GENSCAN and SGP2 predictions to it. Figure 3 shows summaries of the accuracy of SGP2 at the chromosome level in the human and mouse genomes. When compared against ENSEMBL, SGP2 is more accurate than GENSCAN.GENSCAN. It is more specific at the nucleotide level: the average SGP2 specificity is 0.60 for human and 0.61 for mouse, whereas these values for GENSCAN are 0.43 and 0.44. SGP2 is also equally sensitive at the nucleotide level: The average SGP2 sensitivity is 0.82 for human and 0.85 for mouse; these values for GENSCAN are 0.82 and 0.84. Overall, the average SGP2 CCs are 0.70 for human and 0.72 for mouse, and for GENSCAN, the respective averages are 0.59 and 0.61. The accuracy of the SGP2 predictions, moreover, appears to be more consistent across chromosomes than that of the GENSCAN predictions. Interestingly, human chromosome Y is an outlier, with genes in this chromosome being poorly predicted. Genes in chromosome Y appear to be more difficult to predict than genes in other chromosomes for pure ab initio gene prediction programs, because chromosome Y is also an outlier for GENSCAN. SGP2 suffers, in addition, on human chromosome Y because the mouse chromosome Y has yet to be sequenced, and thus there was no comparative information available.

Overall, 23,913 of the human predictions and 24,203 of the mouse predictions overlapped ENSEMBL genes, whereas 95% of the mouse and 93% of the human ENSEMBL genes were among the genes predicted by SGP2. Of the remaining putative novel 20,570 mouse SGP2 genes and 20,193 human SGP2 genes, 10,456 mouse and 9,006 human predictions were found to be similar at $P < 10^{-6}$ to a prediction in the counterpart genome. Of these, 5,960 and 4,909 have multiple exons and are longer than 300 bp. A significant fraction of these putative homologous predictions are likely to correspond to real genes (Guigó et al. 2003). The predictions are interactively accessible through the USCS genome browser (http://genome.cse.ucsc.edu/) and through the DAS server at ENSEMBL (http://www.ensembl.org, under "DAS sources"). The complete set of prediction files is available at http://www1.imim.es/genepredictions/.

Speeding Up TBLASTX Searches

Using TBLASTX to compare human and mouse whole-genome sequences, even in masked form, is quite expensive computationally because of the 6-frame translation on both query and target. To substantially reduce the search time, we used a word size of 5 and sacrificed some sensitivity (see the section above titled "Accelerating TBLASTX Searches" for details). We also penalized stop codons heavily and did not permit gaps. The computation took an estimated 500 CPU days on a farm of Compaq Alphas.

Accuracy in Tables 1 and 2 was computed using default TBLASTX parameters. Table 5 shows the comparative accuracy of TBLASTX and SGP2 predictions, under the default and the speed-up configuration of TBLASTX parameters on the SCIMOG data set. The sensitivity of speed-up TBLASTX searches drops from 0.89 to 0.72, but specificity increases slightly. SGP2 is more robust, and it compensates for some of the sensitivity lost in the TBLASTX search. Overall accuracy for SGP2, as measured by the CC, drops only from 0.95 to 0.93.

Predictions on human chromosome 22 and the whole human and mouse genomes have been obtained with this speed-up configuration of parameters.

DISCUSSION

We have described the program SGP2 for comparative gene finding, and presented the results of its application to the human and mouse genome sequences. Results in controlled benchmark sequence data sets indicate that, by including in-

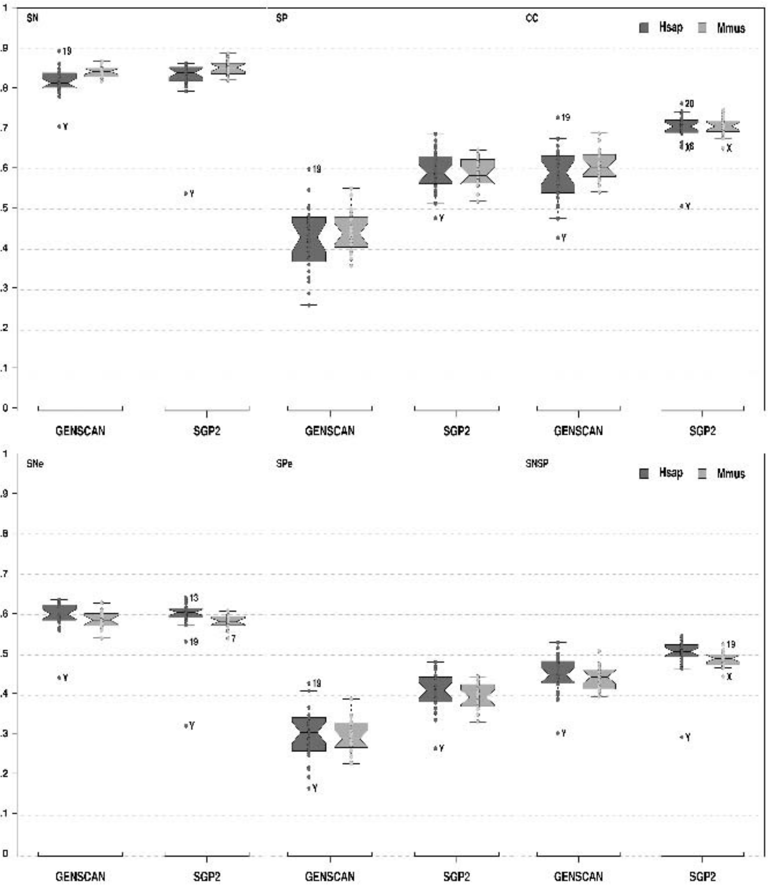


Figure 3 Accuracy of the human and mouse SGP2 and GENSCAN predictions. The accuracy was measured in the entire chromosome sequences using the standard accuracy measures: SN, (sensitivity); SP, (specificity); CC, (correlation coefficient); SNe, (exon sensitivity); SPe, (exon specificity); and SNSP, (average of sensitivity and specificity at exon level). Predictions from both programs were compared against the human and mouse ENSEMBL annotations. Each dot corresponds to the accuracy measure of one chromosome. Chromosome labels are shown for outlier values. The boxplots (Tukey 1977) were obtained using the R-package (http://cran.r-project.org/).

formation from genome sequence conservation, predictions by SGP2 appear to be more accurate than those obtained by pure ab initio programs, exemplified here by GENSCAN and GENEID. Although there is not a significant gain in sensitivity, the specificity of the predictions appears to increase substantially, and a smaller number of false positive exons are predicted.

Indeed, one the major obstacles towards the completion of the catalog of human (mammalian) genes is our inability to assess the reliability of the large number of computational gene predictions that have not been verified experimentally. Whereas the ENSEMBL pipeline produces about 25,000 human and mouse genes, the NCBI annotation pipeline predicts almost 50,000 genes in mouse, and the program GENOMESCAN predicts close to 55,000 genes in this species. Although a large fraction of the ENSEMBL genes correspond to computational predictions without experimental verification, the method is quite conservative, and recent experiments suggest that essentially all ENSEMBL genes are indeed real (Guigó et al. 2003). The problem remains with the tens of thousands of additional computational predictions that are not included in ENSEMBL. A fraction of them are likely to be real, but the question is how large this fraction is. The results obtained here in human chromosome 22 seem to indicate that it may not be very large. Although the existence of hundreds of unidentified genes in this chromosome cannot be completely ruled out, the results strongly suggest that a substantial fraction of these additional computational gene predictions are false positives.

In this regard, the results presented here demonstrate that through the comparison of the human and mouse genomes using SGP2 (or another available comparative gene prediction tool), the false-positive rate can be reduced significantly, and the catalog of mammalian genes better defined. SGP2 predicts a few thousand candidate genes not in ENSEMBL that we believe are worth verifying experimentally. Indeed, the experimental verification of a subset of these provides evidence of at least 1000 previously nonconfirmed genes (Guigó et al. 2003).

The predictions by SGP2 obtained here are, of course, still far from definitively setting this catalog. For one thing, the mouse may be too close a species to human: A large fraction of the sequence has been conserved between the genomes of these two species. Indeed, most sequence conservation between human and mouse does not correspond to coding exons (Mouse Genome Sequencing Consortium 2002), compounding gene prediction. This suggests that the genome of another vertebrate species evolutionarily located between fish and mammals could be of great utility towards closing in the vertebrate (and mammalian) gene catalog.

SGP2 is flexible enough so that it can be easily accommodated to analyze species other than human and mouse. The fact that it can deal with shotgun data at any level of coverage means that as the sequence of a new genome starts becoming available, it can be used to improve the annotation of other already existing genomes. Particularly relevant in this context is a feature of SGP2 (and GENEID) that we have not explored here. SGP2 can produce predictions on top of preexisting annotations. For instance, we could have given to SGP2 the location and exonic coordinates (in GFF format) of known REFSEQ genes (or ENSEMBL), and SGP2 would have predicted genes only outside the boundaries of these genes of

Table 5. Accuracy of TBLASTX and SGP2 Predictions Using "Default" versus Speed-Up Parameters

		Nucleotide			Exon				
		Sn	Sp	CC	Sn	Sp	(Sn+Sp)/2	ME	WE
Default	TBLASTX	0.89	0.76	0.81	—	—	—	0.19	0.11
	SGP2	0.94	0.97	0.95	0.80	0.87	0.83	0.10	0.02
Speed-up	TBLASTX	0.72	0.80	0.75	—	—	—	0.22	0.10
	SGP2	0.88	0.98	0.93	0.77	0.85	0.81	0.12	0.02

already well known exonic structure. Preliminary results indicate that this approach improves gene prediction outside of the preassumed genes, and reduces the rate of chimeric predictions (i.e., predictions encompassing multiple genes). Moreover, we believe that SGP2 can be substantially improved. The flexibility of the SGP2/GENEID framework makes it quite easy to integrate additional information that can contribute to the accuracy of the predictions: synonymous versus nonsynonymous substitution rates in the alignments by TBLASTX, conservation of the splice signals in the informant genome, amino acid substitution matrices specific to the phylogenetic distance between the species compared, etc.

In this regard, the reasons to use the default BLOSUM62 matrix are not obvious. Given the expected sequence similarity between mouse–human orthologs, BLOSUM80 appears to be a better choice. However, we intended to also detect divergent families. Towards that end, the superiority of BLOSUM80 is less clear. We have compared TBLASTX search results on human chromosome 22 against the whole mouse genome. Whereas the HSPs resulting from the BLOSUM62 search cover 84% of the chromosome 22 coding nucleotides, BLOSUM80 HSPs cover 88% of them. However, BLOSUM80 is much less specific than BLOSUM62: 60% of the nucleotides in the BLOSUM62 HSPs fall outside coding regions, compared to 88% for BLOSUM80. It is thus clear that the optimal matrix or combination of matrices for comparative gene-finding using TBLASTX requires further investigation.

Although a large fraction of the human genome sequence has been known for more than a year, the exact number of human genes and their precise definition remain unknown. Gene specification in higher eukaryotic sequences is the result of the complex interplay of sequence signals encoded in the primary DNA sequence, which is only partially understood. Without an exhaustive catalog of human genes, however, the promises of genome research in medicine and technology cannot be completely fulfilled. The work presented here, in which it is shown that human–mouse comparisons can contribute to the completion of the mammalian (human) gene catalog, underscores the importance of the comparisons of the genomes of different organisms to fully understand the phenomenon of life, and in particular to deciphering the mechanism, central to life, by means of which the genome DNA sequence specifies the amino acid sequence of the proteins.

ACKNOWLEDGMENTS

We thank the Mouse Genome Sequencing Consortium for providing the mouse genome sequence as well as support throughout the analysis process. We especially thank Francisco Câmara for arranging the data listed in the gene-prediction page on our group Web site, and for setting up and taking care of our DAS server. We also thank Ian Korf for inspiring discussions regarding the parameters to use in the TBLASTX search. We thank Enrique Blanco, Sergi Castellano, and Moisés Burset for helpful discussions and constant encouragement. This work was supported by a grant from Plan Nacional de I+D (BIO2000-1358-C02-02), Ministerio de Ciencia y Tecnologia (Spain), and from a fellowship to J.F.A. from the Instituto de Salud Carlos III (99/9345).

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D. 1990. Basic local alignment search tool. *J. Mol. Biol.* **215:** 403–410.

Altschul, S.F., Madden, T., Schaffer, A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. 1997. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **25:** 3389–3402.

Bafna, V. and Huson, D.H. 2000. The conserved exon method. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **8:** 3–12.

Batzoglou, S., Pachter, L., Mesirov, J.P., Berger, B., and Lander, E.S. 2000. Human and mouse gene structure: Comparative analysis and application to exon prediction. *Genome Res.* **10:** 950–958.

Birney, E. and Durbin, R. 1997. Dynamite: A flexible code generating language for dynamic programming methods used in sequence comparison. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **5:** 56–64.

Blayo, P., Rouzé, P., and Sagot, M.-F. 2002. Orphan gene finding—An exon assembly approach. *Theoretical Computer Science* (in press).

Borodovsky, M. and McIninch, J. 1993. GenMark: Parallel gene recognition for both DNA strands. *Comput. Chem.* **17:** 123–134.

Burge, C.B. and Karlin, S. 1997. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* **268:** 78–94.

Burge, C.B. and Karlin, S. 1998. Finding the genes in genomic DNA. *Curr. Opin. Struct. Biol.* **8:** 346–354.

Burset, M. and Guigó, R. 1996. Evaluation of gene structure prediction programs. *Genomics* **34:** 353–357.

Claverie, J.-M. 1997. Computational methods for the identification of genes in vertebrate genomic sequences. *Hum. Mol. Genet.* **6:** 1735–1744.

Crollius, H.R., Jaillon, O., Bernot, A., Dasilva, C., Bouneau, L., Fischer, C., Fizames, C., Wincker, P., Brottier, P., Quetier, F., et al. 2000. Estimate of human gene number provided by genome-wide analysis using *Tetraodon nigroviridis* DNA sequence. *Nat. Genet.* **25:** 235–238.

Das, M., Burge, C.B., Park, E., Colinas, J., and Pelletier, J. 2001. Assessment of the total number of human transcription units. *Genomics* **77:** 71–78.

Dunham, I., Hunt, A.R., Collins, J.E., Bruskiewich, R., Beare, D.M., Clamp, M., Smink, L.J., Ainscough, R., Almeida, J.P., Babbage, A., et al. 1999. The DNA sequence of human chromosome 22. *Nature* **402:** 489–495.

Durbin, R., Eddy, S., Crogh, A., and Mitchison, G. 1998. *Biological sequence analysis: Probabilistic models of protein and nucleic acids.* Cambridge University Press, Cambridge.

Gelfand, M.S., Mironov, A.A., and Pevzner, P.A. 1996. Gene recognition via spliced alignment. *Proc. Natl. Acad. Sci.* **93:** 9061–9066.

Gish, W. and States, D. 1993. Identification of protein coding regions by database similarity search. *Nat. Genet.* **3:** 266–272.

Guigó, R. 1998. Assembling genes from predicted exons in linear time with dynamic programming. *J. Comp. Biol.* **5:** 681–702.

Guigó, R. and Wiehe, T. 2003. Gene prediction accuracy in large DNA sequences. In *Frontiers in computational genomic* (eds. M.Y. Galperin and E.V. Koonin), Caister Academic Press, Norfolk, UK.

Guigó, R., Knudsen, S., Drake, N., and Smith, T.F. 1992. Prediction of gene structure. *J. Mol. Biol.* **226:** 141–157.

Guigó, R., Agarwal, P., Abril, J.F., Burset, M., and Fickett, J.W. 2000. Gene prediction accuracy in large DNA sequences. *Genome Res.* **10:** 1631–1642.

Guigó, R., Dermitzakis, E.T., Agarwal, P., Pontig, C.P., Parra, G., Reymond, A., Abril, J.F., Keibler, E., Lyle, R., Ucla, C., et al. 2003. Comparison of mouse and human genomes followed by experimental verification yields an estimated 1,019 additional genes. *Proc. Natl. Acad. Sci.* (in press).

Haussler, D. 1998. Computational genefinding. *Trends in biochemical sciences, supplementary guide to bioinformatics*, pages 12–15.

Jareborg, N., Birney, E., and Durbin, R. 1999. Comparative analysis of noncoding regions of 77 orthologous mouse and human gene pairs. *Genome Res.* **9:** 815–824.

Korf, I., Flicek, P., Duan, D., and Brent, M.R. 2001. Integrating genomic homology into gene structure prediction. *Bioinformatics* **17 Suppl 1:** 140–148.

Meyer, I.M. and Durbin, R. 2002. Comparative ab initio prediction of gene structures using pair HMMs. *Bioinformatics* **18:** 1309–1318.

Miller, W. 2001. Comparison of genomic DNA sequences: Solved and unsolved problems. *Bioinformatics* **17:** 391–397.

Mouse Genome Sequencing Consortium 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420:** 520–562.

Pachter, L., Alexandersson, M., and Cawley, S. 2002. Applications of generalized pair hidden Markov models to alignment and gene finding problems. *J. Comp. Biol.* **9:** 389–400.

Parra, G., Blanco, E., and Guigó, R. 2000. Geneid in *Drosophila*. *Genome Res.* **10:** 511–515.

Pedersen, C. and Scharl, T. 2002. Comparative methods for gene structure prediction in homologous sequences. In *Algorithms in Bioinformatics* (eds. R. Guigó, and D. Gusfield), Springer-Verlag, Berlin, Germany.

Reese, M.G., Hartzell, G., Harris, N.L., Ohler, U., Abril, J.F., and Lewis, S.E. 2000. Genome annotation assessment in *Drosophila melanogaster*. *Genome Res.* **10:** 483–501.

Rinner, O. and Morgenstern, B. 2002. Agenda: Gene prediction by comparative sequence analysis. *In Silico Biol.* **2:** 0018.

Rogic, S., Mackworth, A.K., and Ouellette, F. 2001. Evaluation of gene-finding programs on mammalian sequences. *Genome Res.* **11:** 817–832.

Shoemaker, D.D., Schadt, E.E., Armour, C.D., He, Y.D., Garrett-Engele, P., McDonagh, P.D., Loerch, P.M., Leonardson, A., Lum, P.Y., Cavet, G., et al. 2001. Experimental annotation of the human genome using microarray technology. *Nature* **409:** 922–927.

Tukey, J.W. 1977. *Exploratory data analysis*. pp. 39–41. Addison-Wesley, Boston, MA.

Wiehe, T., Guigó, R., and Miller, W. 2000. Genome sequence comparisons: Hurdles in the fast lane to functional genomics. *Brief. Bioinform.* **1:** 381–388.

Wiehe, T., Gebauer-Jung, S., Mitchell-Olds, T., and Guigó, R. 2001. SGP-1: Prediction and validation of homologous genes based on sequence alignments. *Genome Res.* **11:** 1574–1583.

Yeh, R., Lim, L., and Burge, C. 2001. Computational inference of homologous gene structures in the human genome. *Genome Res.* **11:** 803–816.

Zhang, M.Q. 2002. Computational prediction of eukaryotic protein-coding genes. *Nat. Rev. Genet.* **3:** 698–709.

WEB SITE REFERENCES

http://www.sanger.ac.uk/Software/formats/GFF/; GFF format description page.

http://genome.cse.ucsc.edu/goldenPath/22dec2001/; Human genome sequence goldenpath from Dec. 22, 2001 (hg10) equivalent to NCBI28 build.

http://genome.cse.ucsc.edu/goldenPath/mmFeb2002/; Mouse genome sequence goldenpath from Feb. 2002 (mm2) equivalent to MGSCv3.

http://www.cs.columbia.edu/~vic/sanger2gbd; Victoria Haghighi, Human chromosome 22 curated annotations.

ftp://ftp.ncbi.nih.gov/genomes/H_sapiens/build28_chr_genomescan.gtf.gz; Genomescan predictions from NCBI.

http://genome.cse.ucsc.edu/goldenPath/mmFeb2002/database/ensGene.txt.gz; Mouse ENSEMBL annotations file.

http://blast.wustl.edu; Washington University BLAST Archives

http://genome.cse.ucsc.edu/goldenPath/22dec2001/database/ensGene.txt.gz; Human ENSEMBL annotations file.

http://genome.cse.ucsc.edu; UCSC genome browser.

http://www.ensembl.org; ENSEMBL genome browser.

http://www1.imim.es/genepredictions/; GENEID and SGP2 full data predictions.

http://www1.imim.es/software/sgp2/; SGP2 home page.

http://www1.imim.es/datasets/sgp2002/; SGP2 training data sets page.

Received November 4, 2002; accepted in revised form November 15, 2002.

3.2.2 IMGSC, *Nature*, 420(6915):520–562, 2002

PubMed Accession:

http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&list_uids=12466850&dopt=Abstract

Journal Abstract:

http://www.nature.com/cgi-taf/DynaPage.taf?file=/nature/journal/v420/n6915/abs/nature01262_fs.html

Supplementary Materials:

http://www.nature.com/nature/journal/v420/n6915/suppinfo/nature01262.html
http://genome.imim.es/datasets/mouse2002/

NOTE: Because of copyright restrictions, we cannot offer the article, please follow links for fulltext.

3.3 Validation of Results from Gene Predictors

Annotations from computational gene-finding can be seen as hypotheses about given loci in a genomic sequence encoding cellular functions. Therefore, we initially need to test one of such tools against a controlled data set of reliable annotations to determine its performance. On the other hand, evaluation of predicted genes will be part of the parameters estimation for such software. An iterative procedure may test different program settings under a fixed control set of training sequences in order to determine the parameters that give the best results.

3.3.1 Measures of gene prediction accuracy

To evaluate the accuracy of a gene prediction program, the gene structure predicted by the program is compared with the structure of the actual gene encoded in the problem sequence. As extensively discussed in Burset and Guigó [1996], the accuracy can be evaluated at three different levels of resolution: the nucleotide, exon, and gene levels. These levels offer complementary views of the accuracy of the program. At each level, there are two basic measures: sensitivity and specificity. Briefly, sensitivity (Sn) is the proportion of real elements (coding nucleotides, exons or genes) that have been correctly predicted, while specificity (Sp) is the proportion of predicted elements that are correct. More specifically, if true positive (TP) is the total number of coding elements correctly predicted; true negative (TN), the number of correctly predicted non-coding elements; false positive (FP) the number of non-coding elements predicted as coding; and false negative (FN) the number of coding elements predicted as non-coding. Then, in the gene finding literature, Sn is defined as:

$$Sn = \frac{TP}{TP + FN} \quad ,$$

and Sp as:

$$Sp = \frac{TP}{TP + FP} \quad .$$

Both Sn and Sp take values from 0 to 1, with perfect prediction when both measures are equal to 1. Neither Sn nor Sp alone constitute good measures of global accuracy, since high sensitivity can be reached with low specificity and vice versa. It is desirable to use a single measure for accuracy. In gene finding literature, the preferred such measure at the nucleotide level is the Correlation Coefficient (CC), which is defined as:

$$CC = \frac{(TP \times TN) - (FN \times FP)}{\sqrt{(TP + FN) \times (TN + FP) \times (TP + FP) \times (TN + FN)}} \quad ,$$

and ranges from -1 to 1, with 1 corresponding to a perfect prediction, and -1 to a prediction in which each coding nucleotide is predicted as non-coding and vice versa.

At exon level, these measures determine if predictions correspond to real exons, with the exon boundaries perfectly predicted. The prediction is considered incorrect if only a single base does not correspond to the coordinates of the real exon. Therefore, Sn at exon level measures the proportion of actual exons that have been correctly predicted, and Sp measures the proportion of predicted exons that correspond to actual exons. The average exon prediction accuracy $SnSp$ is computed as:

$$SnSp = \frac{Sn + Sp}{2} \quad .$$

Apart from Sn, Sp and $SnSp$, two extra measures are used to determine the accuracy at exon level: the missed exons (ME) and the wrong exons (WE). ME measures how frequently a predictor completely failed to identify exons (no prediction overlap at all) whereas WE identifies the ratio of exons that do not overlap with any exon of the training data set.

At gene level Sn and Sp measure if a predictor is able to correctly identify and assemble all of the exons of a gene. For a prediction to be counted as TP, all coding exons must be identified, every intron-exon boundary must match exactly, and all the exons must be included in the right gene. In addition, missed genes (MG) and wrong genes (WG) can also be computed in the same way as at the exon level.

The exon level scores discussed above measure how well a predictor recognizes exons and gets their boundaries exactly correct. The gene level scores measure how well a predictor can recognize exons and assemble them into complete genes. In general, gene finders predict the initial and terminal exons very poorly. This often leads to so-called chimeric predictions—one predicted gene encompassing more than one real gene—or to split predictions—where one real gene split in multiple predicted genes. Reese *et al.* [2000] developed two measures to account for these tendencies: split genes (SG) and joined genes (JG). SG is the total number of predicted genes overlapping real genes divided by the number of genes that were split. Similarly, JG is the total number of real genes that overlap predicted genes divided by the number of predicted genes that were joined. A score of 1 is perfect and means that each of the genes from the real genes set overlaps exactly one gene from the set of predicted genes.

3.3.2 Evaluating computational gene-finding results

The evaluations by Burset and Guigó [1996], Rogic *et al.* [2001], and others suffered from the same limitation: gene finders were tested in controlled data sets made of short genomic sequences encoding a single gene with a simple gene structure. These datasets are not representative of the genome sequences that are currently being produced: large sequences of low coding density, encoding several genes and/or incomplete genes, with complex gene structures. This was addressed in the acompanying research article in section 3.3.3, page 54. Table 3.2 on page 56 (Table 1 on page 1632 of Guigó *et al.* 2000) summarizes the results of different gene finding tools in a set of single gene sequences.

The Genome Annotation Assessment Project (GASP) was the first attempt to test the available gene-prediction tools with a well annotated genomic sequence. The 2.9 Mb *Adh* region from *Drosophila melanogaster* was chosen to provide both curated training datasets for the programs and a set of curated annotations to evaluate predictions with them. Table 3.4 on page 79 (Table 3 on page 494 of Reese *et al.* 2000) sums up the results of the gene-finding tools that were evaluated in this experiment.

Table 3.1 on page 27 (Table 4 on page 114 of Parra *et al.* 2003) reports the accuracy of gene-finding programs, including `geneid` and `SGP2`, on human chromosome 22. For the human and mouse comparative analysis we ended up with lots of tables taking into account results for each chromosome sequence and each program, and the evaluations were made with different reference annotation sets. The box-plots shown on Figure 3.5, page 28 (Figure 3 on page 115 of Parra *et al.* 2003), illustrate the differences between gene-finding tools better. This graphical representation provides a compact summary of the different measures being compared, but also shows the dispersion distribution of the data and the outliers. One of the most interesting outliers in the human-mouse analysis was chromosome Y, for which the comparative gene-finding approaches were yielding results similar to those of the "*ab initio*" tools. Of course, this was a result of the lack of orthologous sequences between human and mouse, because for the rodent only female DNA samples were used for sequencing.

In Guigó *et al.* [2003], a protocol for selecting computational predictions to be tested by experimental means, via RT-PCR in this case, is described. `SGP2` results from the gene prediction pipeline, detailed in section 3.2, were classified into three groups in function of the homolgy between the human and mouse predictions and the conservation of their exonic structures. Table 3.5 on page 92 (Table 1 on page 1143 of Guigó *et al.* 2003) sumarizes the RT-PCR success rate within each of those groups. Figure 3.6 on page 39 (Figure 16 on page 540 of Waterston *et al.* 2002) shows the structures, side by side, of a human and mouse predicted new homologue of *dystrophin*, for which an exon pair from the mouse gene was verified by RT-PCR. Another example, a novel homolog to *Drosophila melanogaster* brain-specific homeobox protein, can be found on Figure 3.8, page 91, for which the primers and RT-PCR results are depicted on the same page in Figure 3.9 (Figures 2 and 3 on page 1142 respectively, of Guigó *et al.* 2003). A database was built for the 476 gene structures that were tested by RT-PCR. It contains not only the sequences and coresponding annotations for those genes, but also the results yielded from each RT-PCR test done in 12 different mouse tissues. Figure 3.10 on page 95 shows the web interface we have created for that database.

All results indicate that there is room for improvement in the computational gene prediction field. Efforts to provide more accurate gene-finding tools, as well as more reliable annotations, are ongoing. The best example of such efforts is the ENCODE project [ENCODE Project Consortium, 2004]. During its pilot phase, the procedures that can be applied cost-effectively and at high-throughput to accurately and comprehensively characterize large sequences, will be evaluated.

3.3.3 Guigó *et al*, *Genome Research*, 10(10):1631–1642, 2000

PubMed Accession:

http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&list_uids=11042160&dopt=Abstract

Journal Abstract:

http://www.genome.org/cgi/content/abstract/10/10/1631

Supplementary Materials:

http://genome.imim.es/datasets/gpeval2000/

Methods

An Assessment of Gene Prediction Accuracy in Large DNA Sequences

Roderic Guigó,[1,3] Pankaj Agarwal,[2] Josep F. Abril,[1] Moisés Burset,[1] and James W. Fickett[2]

[1]Grup de Recerca en Informática Mèdica, Institut Municipal d'Investigació Mèdica, Universitat Pompeu Fabra, E-08003 Barcelona, Spain; [2]Department of Bioinformatics, SmithKline Beecham Pharmaceuticals Research and Development, King of Prussia, Pennsylvania 19406, USA

One of the first useful products from the human genome will be a set of predicted genes. Besides its intrinsic scientific interest, the accuracy and completeness of this data set is of considerable importance for human health and medicine. Though progress has been made on computational gene identification in terms of both methods and accuracy evaluation measures, most of the sequence sets in which the programs are tested are short genomic sequences, and there is concern that these accuracy measures may not extrapolate well to larger, more challenging data sets. Given the absence of experimentally verified large genomic data sets, we constructed a semiartificial test set comprising a number of short single-gene genomic sequences with randomly generated intergenic regions. This test set, which should still present an easier problem than real human genomic sequence, mimics the ~200kb long BACs being sequenced. In our experiments with these longer genomic sequences, the accuracy of GENSCAN, one of the most accurate ab initio gene prediction programs, dropped significantly, although its sensitivity remained high. Conversely, the accuracy of similarity-based programs, such as GENEWISE, PROCRUSTES, and BLASTX, was not affected significantly by the presence of random intergenic sequence, but depended on the strength of the similarity to the protein homolog. As expected, the accuracy dropped if the models were built using more distant homologs, and we were able to quantitatively estimate this decline. However, the specificities of these techniques are still rather good even when the similarity is weak, which is a desirable characteristic for driving expensive follow-up experiments. Our experiments suggest that though gene prediction will improve with every new protein that is discovered and through improvements in the current set of tools, we still have a long way to go before we can decipher the precise exonic structure of every gene in the human genome using purely computational methodology.

The nucleotide genomic sequence is the primary product of the Human Genome Project, but a major short- and mid-term interest will be the amino acid sequences of the proteins encoded in the genome. Thus, methods that reliably predict the genes encoded in genomic sequence are essential, and computational gene identification continues to be an active field of research (for reviews, see Fickett 1996; Claverie 1997; Guigó 1997a; Burge and Karlin 1998; Haussler 1998). A new generation of gene prediction programs based on Hidden Markov Models (Burge and Karlin 1997) have shown significantly greater accuracy than previous programs based on other methodologies (Burset and Guigó 1996). Conversely, as the databases of known coding sequences increase in size, gene prediction methods based on sequence similarity to coding sequences, mainly proteins and ESTs, are becoming increasingly useful and are routinely used to identify putative genes in genomic sequences (The *C. elegans* Sequencing Consortium 1998). We have recently published an evaluation of sequence similarity-based gene prediction methods, in particular of EST-based gene prediction (Guigó et al. 2000). The accuracy of gene identification programs, however, has usually been estimated on controlled data sets made of short genomic sequences encoding a single and complete gene with a simple structure. Moreover, these data sets are often similar if not overlapping, to the sets of sequences on which the programs have been trained. Thus, these data sets are not representative of the sequences being produced at the genome centers, which are mostly large sequences of low coding density, encoding several genes or incomplete genes with complex gene structure. It is thus difficult to know how well the figures of accuracy estimated in the controlled benchmark data sets extrapolate to actual genomic sequences. Furthermore, programs that combine both sequence similarity and ab initio gene finding approaches, and those that predict genes by producing a splicing alignment between a genomic sequence and a candidate amino acid sequence have become recently available, such as PROCRUSTES (Gelfand et al. 1996) and GENEWISE (Birney and Durbin 1997), (http://www.sanger.ac.uk/Software/Wise2/). Programs that align genomic sequences with

[3] **Corresponding author.**
E-MAIL rguigo@imim.es; FAX 3493-221-3237.

Article and publication are at www.genome.org/cgi/doi/10.1101/gr.122800.

EST sequences, such as EST__GENOME (Mott 1997), could also be included in this category. These programs promise highly accurate predictions, but at the cost of greater computational time. However, this increase in accuracy has not been well-quantified on challenging data sets. The effects of the degree of similarity between the candidate homolog and the genomic sequence also deserve careful evaluation.

We believe a more realistic evaluation of the currently available gene prediction tools on challenging data sets would be useful. Ideally, one would like to benchmark the computational gene identification programs in real genomic sequences. The main problem is that most real sequences the structure of the genes has not been verified exhaustively by experimental means, and thus it is impossible to calibrate the accuracy of the predictions. Only recently, extensively annotated large genomic sequences from higher eukaryotic organisms have become available from the human genome (http://www.hgmp.mrc.ac.uk/Genesafe) and from the fly genome (http://www.fruitfly.org/GASP1/). In spite of the experimental analysis, the possibility of undetected genes in the sequence cannot be easily ruled out, which makes accuracy difficult to measure. Here, we attempt to overcome the lack of well-annotated large genomic sequences by constructing semiartificial ones. In these semiartificial sequences, known genomic sequences have been embedded in simulated intergenic DNA, and therefore, the location of all coding exons is known. Although the approach may seem unrealistic, we believe that the results obtained are instructive with regard to the accuracy of currently available gene identification tools.

We evaluate the accuracy of representatives of a wide variety of computational gene identification approaches: GENSCAN (Burge and Karlin 1997), an ab initio genefinder; BLASTX (Altschul et al. 1990; Gish and States 1993), a genefinding-oriented similarity search program; and PROCRUSTES (Gelfand et al. 1996) and GENEWISE (Birney and Durbin 1997), genefinders based on aligning a genomic DNA sequence fragment to a homologous protein sequence. We evaluate these programs on two benchmark data sets: A set of well-annotated single-gene DNA sequences, and a set of semiartificial genomic (SAG) sequences created by embedding the single-gene sequences from the first data set in simulated intergenic DNA.

RESULTS

We investigated the accuracy of the gene prediction tools (GENSCAN, PROCRUSTES, GENEWISE, BLASTX) described in Methods on two benchmark sets. In all cases, sequences were masked previously for repeated regions using REPEATMASKER (A. Smit and P. Green, unpubl.). The gene predictions obtained using the different tools were compared with the actual gene annotations using the accuracy measures described Methods.

Accuracy in Single Gene Sequences

Table 1 shows the accuracy of the different gene prediction tools on h178, the set of single gene sequences.

GENSCAN's accuracy is comparable to that reported earlier (Burge and Karlin 1997). On average, 90% of the coding nucleotides and 70% of the exons are predicted correctly by GENSCAN. Only 7% of the actual exons are missed completely, and only 9% of the predicted exons are wrong. We believe this is close to the maximum accuracy that can be achieved using currently available ab initio gene prediction programs.

The quality of the gene models inferred from BLASTX searches depends on the strategy used. Default usage of BLASTX produced poorer predictions than more sophisticated strategies. (Results for BLASTX default correspond to those published in Guigó et al. 2000.) Discrepancies between numbers in Table 1 and those reported in Guiqoet al. (2000) are due to the differences in the way the accuracy measures are summarized. In Guigó et al. 2000, we computed the accuracy measures on each test sequence, and averaged all of them. Here, we compute the accuracy measures globally from the total number of prediction successes and failures (at the base or exon level) on all sequences. The default BLASTX strategy produces reasonably high sensitivity (0.91) by projecting all HSPs over a given threshold along the query DNA sequence, but the sensitivity rises to an amazing 0.97, if the topcomboN fea-

Table 1. Accuracy of Gene Prediction Tools in the Set of Single Gene Sequences (h178)

		Nucleotide			Exon				
Program	No.	Sn	Sp	CC	Sn	Sp	$\frac{Sn + Sp}{2}$	ME	WE
GenScan	177	0.93	0.90	0.90	0.78	0.75	0.76	0.08	0.10
Blastx default	175	0.91	0.79	0.82	0.04	0.04	0.04	0.12	0.05
Blastx topcomboN	174	0.97	0.80	0.86	0.04	0.04	0.04	0.08	0.05
Blastx 2 stages	175	0.90	0.92	0.90	0.10	0.12	0.11	0.19	0.02
GeneWise	177	0.98	0.98	0.97	0.88	0.91	0.89	0.06	0.02
Procrustes	177	0.93	0.95	0.93	0.76	0.82	0.79	0.11	0.04

ture is used. The topcomboN feature eliminates the need for low-complexity filters (seg + xnu), and for strict secondary HSP cutoff (S2 threshold). Surprisingly, its use does not appear to hurt specificity. The two-stage method (in which the top homolog with low-complexity filtering is chosen to build the BLASTX model with topcomboN in the second stage) increases specificity from 0.79 to 0.92. Using a single protein to build a model improves specificity because the noise from the less significant hits is reduced. But the two stage method does have lower sensitivity from a lack of information from the weaker secondary hits. However, this is still the best purely BLASTX-based strategy in terms of either specificity or overall accuracy, and the numbers are comparable to the accuracy of ab initio gene finders at the nucleotide level.

The proteins encoded by the sequences in h178 are mostly included in the nonredundant database of amino acid sequences (*nr*). However, BLASTX still does not produce perfect predictions. This certainly has an artefactual component: We have discovered a few annotation errors in h178. However, perfect gene predictions from BLASTX searches are intrinsically impossible because of the inability of BLASTX to predict the splice boundaries when they occur within codons (this especially affects its accuracy at the exon level, which is actually rather meaningless for BLASTX). In this regard, splicing alignment or sequence similarity-based gene prediction tools (SSBGP), such as GENEWISE and PROCRUSTES could, in principle, result in more accurate predictions. Thus, the protein sequence with the lowest *P* value after the BLASTX search was given to PROCRUSTES and GENEWISE to model their gene predictions. SSBGP tools improved the accuracy of the gene predictions inferred directly from BLASTX searches, and also slightly outperform GENSCAN in this set. GENEWISE predictions with an overall accuracy of 0.97, in particular, were close to perfect given the intrinsic inaccuracy of the database annotation considered to be the gold standard here. Of course, there is a price paid in computational time, and GENEWISE is expensive with its linear-memory dynamic programming technique.

GENSCAN accuracy, in theory, should be unaffected, whether the query sequence encodes genes for which a close homolog, remote homolog, or no homolog exists. GENEWISE and PROCRUSTES accuracy, on the other hand, should decrease as the homology becomes distant, and these programs have little utility if a homolog does not exist.

As we have already pointed out, *nr* database contains protein translations of most of the genes in our data set, which could be a significant drawback of the data set. It is difficult (if not impossible) to come up with criteria for eliminating just the translations. Mouse orthologs are often 100% identical at the protein level and variants of the same protein may be highly (98%–99%) identical. Thus, we chose to evaluate the effect of the similarity level (*P* value) of an available homolog on the accuracy of GENEWISE and PROCRUSTES by considering a variety of *P* value bins. Conceptually, identical or close to identical proteins would fall in the most significant *P* value bin, and other bins would be devoid of identical hits.

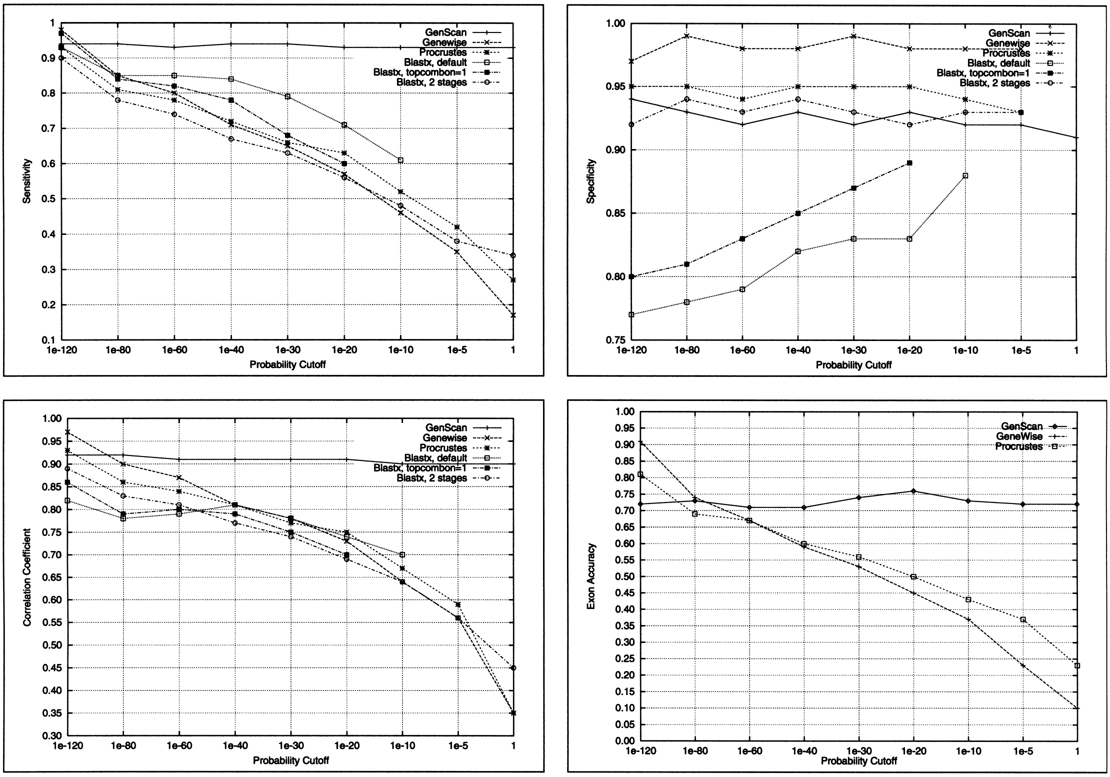
A set of Blast-probability (*P* value) thresholds was chosen to provide bins with varying levels of similarity (10^{-120}, 10^{-80}, 10^{-60}, 10^{-40}, 10^{-30}, 10^{-20}, 10^{-10}, and 10^{-5}). For each of these *P* values (10^{-80}, for instance), we performed the following experiment. After running BLASTX against *nr* for the DNA sequences in h178, we discarded for each DNA sequence all HSPS corresponding to all protein sequences with a *P* value below cutoff (as if we were ignoring all known amino acid sequences over a given level of similarity to the protein encoded in the query DNA sequence). Then, the protein with the remaining top hit below the next higher *P* value threshold (10^{-60}, in the case of the example) was used, if it existed, as a candidate homolog for the SSBGP tools. If there was no protein hit in the bin (10^{-80} to 10^{-60} in the example) then this gene was discarded for the evaluation of this bin.

Thus, the BLASTX gene models are based on all the protein homologs with probability higher than the threshold considered. The *P* value thresholds were chosen so as to generate roughly equal numbers of data points (sequences from h178) for each set. The minimum number of data points in any set is 73, large enough to avoid significant sampling bias.

The accuracy results as a function of *P* value of the homologs are shown in Figure 1. GENSCAN performance is expected to be constant, and was for the most part; the minor variations are because of changes in the data set. Only a fraction of the genes had homologs in each of the bins, thus the data set changed a little from bin to bin. The overall performance of SSBGP tools suffered substantially as the similarity decreased. Somewhat surprisingly, the performance of GENSCAN is superior to that of SSBGP tools even at rather high levels of similarity (*P* value between 10^{-80} and 10^{-60}). When the similarity is strong, GENEWISE appears to outperform PROCRUSTES in the h178 sequence set. However, when the similarity is weak the difference in performance between the two tools at the nucleotide level is small, and for low levels of similarity PROCRUSTES seems to outperform GENEWISE, particularly at the exon level. This is not unexpected considering the design of these programs: GENEWISE is primarily a sequence alignment tool, and thus it performs very well when there is strong sequence similarity. PROCRUSTES is more of a gene prediction program; it possibly encodes a more sophisticated splice site and exon model, which allows for better exon prediction at low levels of similarity. As shown in Figure 3, a decrease in accuracy for sequence similarity-based methods is most likely a result of the decline in sensitivity, while specificity remains high, which is a very desirable feature.

Interestingly, when the similarity is weak (P value > 10^{-20}), the advantage of sophisticated SSBGP tools as opposed to direct gene modeling from database searches such as those performed by BLASTX, seems to vanish. It is not unlikely that when the similarity is weak, the query DNA sequence and the top database search homolog share only a conserved domain. In such cases, SSBGP, relying on sequence similarity only to the top homolog, are only able to detect the part of the gene exonic structure encoding these domains. Direct gene modeling from BLASTX search results builds on all potential homologs (not only the top one); thus, weak homologs that share different conserved regions with the gene encoded in the DNA sequence may allow for better recovery of the overall exonic structure of the gene. In fairness to GENEWISE and PROCRUSTES, they can be used with multiple protein homologs and complete gene models synthesized, but that is computationally expensive and analytically problematic. Figure 1 illustrates an extreme example. A possible solution (at least when using GENEWISE) is to build a profile or an HMM based on the top few homologs and then align this profile with the target genomic sequence.

Conversely, when the similarity with the top ho-



P-value cutoff	1e-120	1e-80	1e-60	1e-40	1e-30	1e-20	1e-10	1e-05	1
# sequences	73	99	100	116	102	108	109	96	119
P-value log-average	1e-206	1e-104	1e-73	1e-54	1e-37	1e-27	1e-17	1e-8	1e-3

Figure 1 The accuracy of the gene prediction tools as a function of the similarity to the chosen homolog. For each P-value cutoff, the homolog with the lowest P value above the cutoff was chosen to build the gene prediction models. The table indicates the different ranges considered, the log-average of the P values in each range, and the number of sequences with acceptable homologs in the range. For example, there were 99 sequences in h178 for which after discarding all hits with P value < 10^{-120}, the top remaining hit had a P value < 10^{-80}. There were 73 sequences for which the top hit had a P value < 10^{-120}, and 119 sequences for which the top hit had a P value > 10^{-5}.

molog is weak, the BLASTX search picks up only the stronger regions of similarity between the homolog and the gene encoded in the query sequence, although lower levels of sequence similarity are shared in other regions between the protein and the query DNA sequences. These can be detected by the SSBGP tools (Fig. 1). Finally, in other cases, both situations occur simultaneously, and direct gene modeling from BLASTX search and SSBGP tools may complement each other to produce a more accurate overall prediction (Fig. 1).

Examining the data in Table 1 and Figure 1, one may be tempted to conclude that the gene identification problem is almost solved. When a strong homolog exists, programs like GENEWISE and PROCRUSTES are likely to pick up the correct exon structure; when such a homolog does not exist, programs like GENSCAN will still be able to recover most of this structure. This, we believe, is rather optimistic, as the sequence set in which these programs have been tested is extremely easy. Although the results obtained are instructive of the comparative performance of the tools, they cannot necessarily be extrapolated to the performance of these tools in the large genomic sequences. In the next section, we present the results obtained on evaluating the tools on a set of simulated genomic sequences, which we believe provide a more realistic estimation of the actual accuracy of the gene prediction tools in large genomic sequences.

Accuracy in Semiartificial Genomic Sequences

A SAG data set containing known genes in random intergenic context (as described in Methods) was constructed to check if the accuracy measures from the previous section extrapolate to larger, more difficult data sets.

Because each SAG sequence contains multiple genes, the choice of the set of protein homologs to predict all the genes was no longer trivial. For ease of evaluation, we used the knowledge of the genes to pick these homologs, but there are other techniques that can be used to pick up a single candidate homolog for each gene-like region. In short, the top-scoring protein homolog from the BLASTX search for each of the genic sequences was used by GENEWISE and PROCRUSTES to predict the gene based on sequence similarity. For instance, artificial sequence AGS01 was obtained by embedding EMBL sequences HS10116, HSDNAAMHI, and HSNUCLEO in artificial intergenic DNA, with BLASTX top homologs being NCBI:gi 134635, 1136442, and 128841, respectively. The GENEWISE and PROCRUSTES predictions on the artificial sequence AGS01 were obtained by three independent executions of the programs, with each of the above top homolog proteins in turn. The programs were executed to predict genes on both strands and the model on the strand with the higher score was used to assess accuracy. This approach isolated the issue of the accuracy of these programs if the genomic sequence is large and the gene is encoded only in a small region of this sequence. There are other factors, such as the ability to choose the correct set of homologs that affect accuracy, but these factors were similar for all the programs, and other suboptimal (but perhaps more realistic) techniques would lead to lower accuracy. Thus, the accuracy numbers for the semiartificial sequences are not underestimated.

Table 2 shows the accuracy of the gene identification tools in Gen178, the set of simulated genomic sequences. As expected from theoretical considerations, SSBGP tools were mostly unaffected by the inclusion of genic sequences in the random intergenic-like DNA. PROCRUSTES appears to be less robust than GENEWISE when analyzing large genomic sequences. In particular, there is a significant decrease in specificity at the exon level (from 0.82 to 0.75), the likely result of predicting a relatively large number of small exons in otherwise noncoding DNA [wrong exons (WE) increasing from 0.04 to 0.16]. The comparatively low decrease in specificity at the nucleotide level, from 0.95 to 0.94, suggests that most of these false exons are rather short. Surprisingly, PROCRUSTES sensitivity at

Table 2. Accuracy of Gene Prediction Tools in the Set of Semiartificial Genomic (SAG) Sequences (Gen178)

					Exon						
		Nucleotide					$\frac{Sn + Sp}{2}$			Gene	
Program	No.	Sn	Sp	CC	Sn	Sp		ME	WE	MG	WG
GenScan	43	0.89	0.64	0.76	0.64	0.44	0.54	0.14	0.41	0.03	0.28
		0.92	*0.92*	*0.91*	*0.76*	*0.76*	*0.76*	*0.09*	*0.09*		
GeneWise	43	0.98	0.98	0.97	0.88	0.91	0.89	0.06	0.02		
		0.98	*0.98*	*0.97*	*0.88*	*0.91*	*0.89*	*0.06*	*0.02*		
Procrustes	43	0.93	0.94	0.93	0.80	0.75	0.77	0.10	0.16		
		0.93	*0.95*	*0.93*	*0.76*	*0.82*	*0.79*	*0.11*	*0.04*		

(Italics) The accuracy values in the set of single gene sequences (from Table 1).

the exon level is slightly higher in the set of artificial sequences than in the set of single gene sequences.

The accuracy of BLASTX was not affected by the intergenic context (data not shown) because no hits with a P value more significant than 10^{-10} were found in the simulated DNA.

Accuracy of ab initio gene finders suffered substantially in the set of artificial genomic sequences. Because of the tendency of gene finders to overpredict exons, one would expect that by placing the genic sequences in the simulated-intergenic context, some loss of specificity would be observed, with programs predicting perhaps a few extra exons in otherwise random DNA. On the other hand, one would expect the sensitivity to remain essentially constant as the exons predicted in the genic sequences should still be predicted when these are included in simulated-intergenic DNA. However, a significant decrease in specificity is observed (Table 2). For instance, GENSCAN specificity at the exon level drops to 0.64 from 0.92, and the proportion of WEs climbs to 41% from 9% in the single gene sequences. In addition, a significant decrease in sensitivity is also observed, with programs failing to predict exons that were correctly identified in the single gene sequences. For instance, the proportion of missing exons increases for GENSCAN from 9% to 14%. Almost 30% of the GENSCAN genes are predicted in the simulated-intergenic DNA. For ab initio gene finders, we believe these accuracy values (on SAG sequences) are more representative of their true accuracy on large genomic sequences than those obtained in the typical single gene benchmark experiments.

Figure 2 shows the predictions of the different programs in one of the artificially generated genomic sequences (~157-kb long). As mentioned, SSBGPs predict the genic structure of the artificial genomic sequence rather well. Performance of ab initio gene finders, on the other hand, degrades substantially.

Although all genes predicted by GENSCAN overlap real genes, it still predicts a large number of false positive exons. In addition, even when predicting the exons correctly, their assembly into genes is often incorrect. For instance, in the sequence in Figure 2, GENSCAN has difficulty in predicting the correct gene boundaries, and it expands the gene beyond its actual limits. In the lower portion of the Figure 2, we compare the predictions in the region between positions 23,000 and 41,000 from the SAG sequence to the predictions on just the actual genic sequence (without the random context). GENSCAN performance suffers substantially from this inclusion in pseudointergenic context. One explanation is that GENSCAN uses the wrong isochore model for this sequence: the actual isochore structure being destroyed by the usage of artificial intergenic context. In such a case, decrease in performance would be an artifact of our SAG sequences rather than a feature of GENSCAN. Experiments with gene finders other than GENSCAN (data not shown) indicate that such a decrease in performance is not specific to GENSCAN, but rather a general feature of ab initio gene finders.

As with the set of single gene sequences, the comparison of GENSCAN with SSBGP tools is not strictly fair. The SSBGPs are affected by the existence of closer homologs, while GENSCAN is not affected. To study the effects of the range of similarity on the accuracy of gene prediction in the SAG data set, we extracted two different sets of SAG sequences. In the first set, each gene in each SAG sequence has a strong homolog (BLASTX P value $< 10^{-50}$), and in the other set, each gene in each sequence had a moderate homolog (BLASTX P value between 10^{-50} and 10^{-6}). Some of the genes in the second set also had better homologs which were ignored for this analysis. The results are shown in Table 3. If the similarity is strong, the sequence similarity-based methods perform very well, outperforming ab initio tools (as in Table 2). However, if the average similarity between the genes encoded and the known proteins is only moderate (though perhaps, still better than expected for real genomic sequences), the performance of these tools is similar to the performance of GENSCAN. At the exon level, the overall accuracy stays at ~50%. A very similar accuracy has also been observed independently on test sets on actual genomic sequences (http://predict.sanger.ac.uk/th/brca2/; see Discussion). We believe this is still an overestimation of the actual accuracy of these tools in real genomic sequences.

DISCUSSION

Computational genefinders produce acceptable predictions of the exonic structure of the genes when analyzing single gene sequences with very little flanking intergenic sequence, but are unable to correctly infer the exonic structure of multigene genomic sequences. In particular, ab initio genefinders predict and utilize intergenic boundaries poorly. Conversely, as our results indicate, sequence similarity searches on databases of known coding sequences are extremely helpful in deciphering the exonic structure for the genes that have known homologs. For very strong similarity, SSBGP tools appear to be the most useful. Surprisingly even for genes predicted based on homologs with a moderate degree of similarity ($10^{-50} < P$ value $< 10^{-6}$), GENSCAN performs comparably to SSBGP programs. It appears that at such levels of similarity, potential splice signals and statistical biases in the sequence composition carry information comparable to sequence similarity for the purposes of identifying coding regions. It is possible that the use of SAG sequences does not provide a realistic scenario to test the accuracy of computational gene finders. Ideally, one would like to use large genomic sequences with gene structure experi-

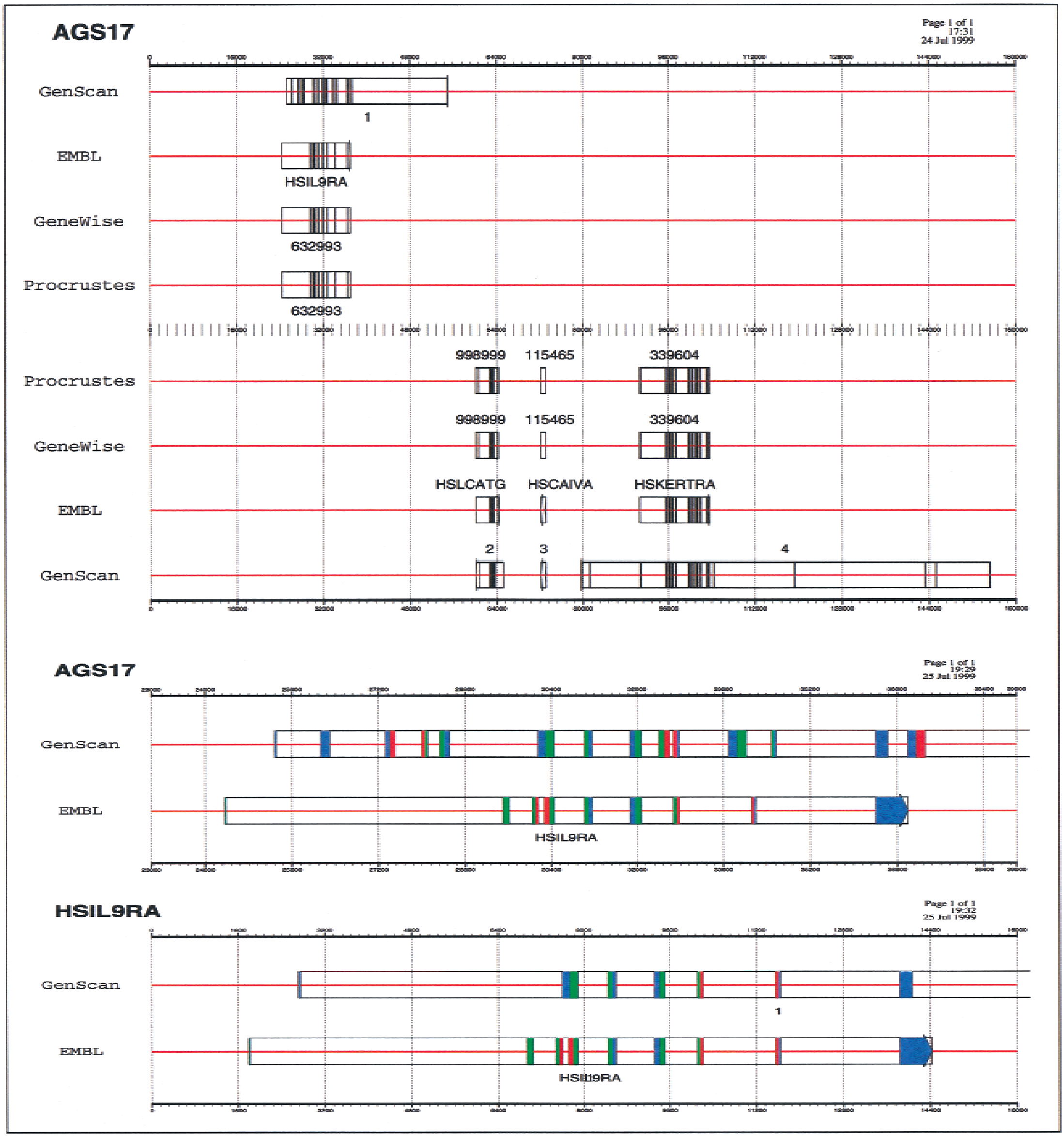
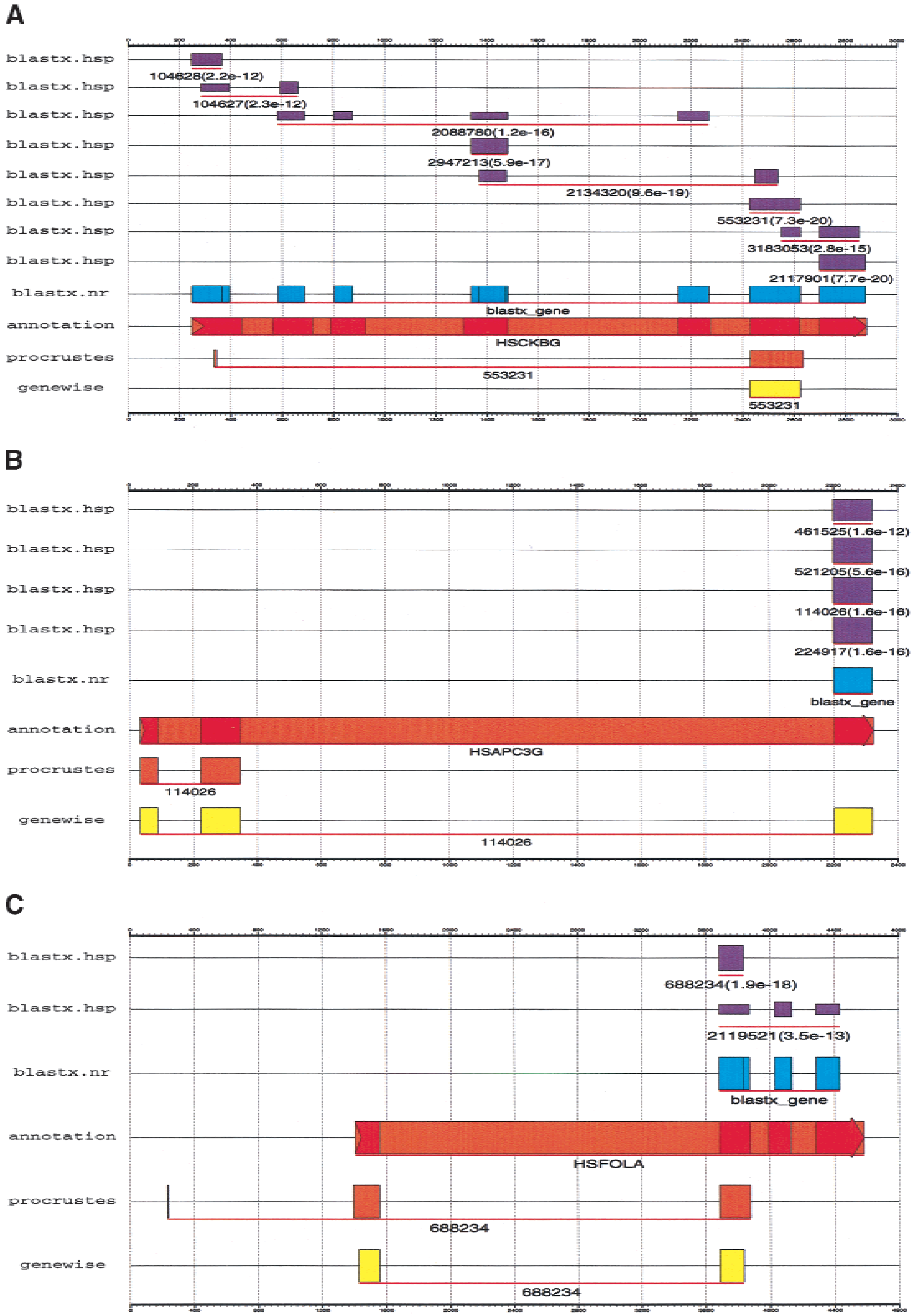


Figure 2 (AGS17, *top*) Gene predictions in one of the artificial genomic sequences. The row EMBL indicates the coordinates of the actual genes. Exons corresponding to the same gene (or predicted to be in the same gene) are linked by a box. (AGS17, *middle*) Predictions of GENSCAN finders in the region 23,000 to 41,000 from the semiartificial genomic sequence. (HSIL9RA, *bottom*) The predictions improve if GENSCAN is provided only the 18,000-bp long genic sequence that has been inserted in this region. This figure, as well as Fig. 1, has been prepared using gff2ps. (Abril and Guigó 2000)

mentally verified. However, experimentally verifying each and every gene along with alternative splice structures in a large genomic sequence remains a difficult challenge. Techniques such as exon-trapping (Church et al. 1994) have high sensitivity but poor specificity, while RT–PCR or identifying a cDNA clone for every transcript can be fairly specific (Hochgeschwender 1992), but have less than perfect sensitivity and are dependent on finding a tissue in a developmental stage under an environmental condition in which that gene (or alternative gene product) is expressed. In particular, proving that a piece of sequence (that appears coding



to gene-prediction programs) is not coding is extremely difficult. Thus, even though there are a number of attempts to consolidate genomic gene prediction data sets [Banbury Cross (http://igs-server.cnrs-mrs.fr/igs/banbury), GeneSafe (http://www.hgmp.mrc.ac.uk/Genesafe), GASP (http://www.fruitfly.org/GASP1/)], the number of experimentally well-annotated large genomic sequences remains small, and even in those

Figure 3 If the candidate protein sequence is a remote homolog, direct gene modeling from BLAST-like database searches may have different predictions compared to more sophisticated SSBGP tools. (*A*) EMBL DNA sequence HSCKBG was compared with the protein sequences in the nr sequence database using BLASTX. Hits with *P* value < 10^{-20} were discarded, the top remaining corresponded to a fragmentary protein sequence gi:553231. Not surprisingly, only a small fraction of the actual gene was recovered using this homolog by either GENEWISE or PROCRUSTES. Other choices of homologs may have yielded different predictions but none of them by themselves appears to be perfect. Conversely, the gene model derived directly from the BLASTX search reproduces the exonic structure of the gene fairly well. Thus, even though upon discarding the close homologs, the remaining proteins individually showed only little overall similarity to the encoded protein product, as a collection they enable to walk its exonic structure. (*B*) If database protein sequences with hits below *P*-value = 10^{-20} are discarded, BLASTX is able to detect significant similarity between only one of the encoded exons in EMBL sequence HSPAC3G and the remaining protein sequences in the database. But with the top homolog among these, the SSBGP tools (GENEWISE in particular) are able to infer the correct exonic structure, picking up both the additional upstream exons. This is because the SSBGP tools are able to detect more distant sequence relationships than BLASTX with our choice of thresholds or because (as in this case) coding exons occur in low-complexity regions, which are usually masked when performing BLASTX searches to avoid large numbers of false positives. (*C*) In another case, direct gene modeling from BLASTX searches and SSBGP tools can complement each other to produce more accurate gene predictions. As in *A* and *B*, HSP hits below *P*-value = 10^{-20} were ignored after comparing EMBL sequence HSFOLA with the nonredundant protein sequence database.

cases, the reliability of the annotation is difficult to assess (Reese et al. 2000). To compensate for the lack of these verified data sets, we have built semiartificial data sets with known genes placed in the context of random intergenic sequence. This ensures that all the genes in these sequences are known. In fact, most of these genes have fairly small genomic spread (i.e., none of the introns is very large), and a number of the ab initio gene prediction programs have been trained on them. This should make this data set easy for most programs. However, our model for intergenic sequence is possibly imperfect for at least two reasons: The genes are not necessarily placed in the correct isochore context; and the apparent codon composition in the simulated intergenic DNA may be different from that of actual intergenic sequence. These imperfections may conceivably make gene prediction more difficult on this data set for ab initio programs, but we think these are more than offset at least in part by the small genes and the fact that the programs have partly trained on these genes. Overall, the sensitivity and specificity numbers are most instructive in the relative context. The sensitivity of most tools remains high even when confronted with large intergenic sequences, but the specificity of the ab initio tools drops because of large intergenic regions.

Interestingly, the accuracy reported here for GENSCAN is very similar to the accuracy found in the *BRCA2* region (Chruch et al. 1994; Couch et al. 1996); probably the best annotated human genomic region from an experimental standpoint. BRCA2 region is a large genomic tract with multiple genes, thus, a difficult data set for most gene prediction programs. At the exon level, Tim Hubbard and Richard Bruskiewich (Sanger Center, UK) report for GENSCAN in this region a sensitivity of 0.63 (termed *coverage* there) and a specificity of 0.38 (termed *accuracy* there) (http://predict.sanger.ac.uk/th/brca2/). As anticipated, these values are slightly worse than the ones we have found here in the SAG data set (0.64 and 0.44, respectively). This seems to indicate that the approach of building artificial genomic sequences is not too unrealistic, and that it could be useful both for training and testing gene prediction programs. Results in these sequences, however, should be taken as an upper bound estimate of the accuracy of the programs in real genomic sequences.

There is a growing class of gene identification programs that combine both sequence similarity and traditional coding potential measures, such as Genie (Kulp et al. 1996 1997), HMMgene (Krogh 1997), and GSA (Huang et al. 1997). Unfortunately, because of a

Table 3. Accuracy of Gene Prediction Tools in the Set of Semiartificial Genomic Sequences, When Either Strongly or Moderately Similar Sequences are Used to Model the Genes

	Strong similarity *P* Value < 10^{-50} 17 SAG sequences						Moderate similarity 10^{-50} < *P* value < 10^{-6} 26 SAG sequences					
	Nucleotide			Exon			Nucleotide			Exon		
Program	Sn	Sp	CC	Sn	Sp	$\frac{Sn + Sp}{2}$	Sn	Sp	CC	Sn	Sp	$\frac{Sn + Sp}{2}$
GenScan	0.91	0.66	0.77	0.67	0.46	0.56	0.91	0.61	0.74	0.67	0.43	0.55
GeneWise	0.99	0.99	0.99	0.90	0.93	0.91	0.68	0.98	0.81	0.46	0.63	0.54
Procrustes	0.92	0.96	0.94	0.80	0.75	0.77	0.66	0.79	0.72	0.48	0.32	0.40

The geometric mean of the *P* values of the strong similarity sequences was 10^{-135} and for the weaker similarity group it was 10^{-39}.

lack of public availability at the time of the initiation of this study, their evaluation will have to await a future analysis.

EST similarity can also provide useful information regarding gene structure for ~85% of the common genes (Guigó et al. 2000). A set of single gene sequences in h178 was used to optimize a method for deriving exonic structures from EST matches. When using the EST sequences in the public databases, the method yielded an accuracy of $Sn = 0.72$, $Sp = 0.87$, and $CC = 0.69$ at the nucleotide level, when predicted gene structures were compared to the annotated mRNA (not the coding) exonic structure. Other secondary questions regarding EST-based gene prediction may also be important, such as the extent to which EST matches help in delineating the gene boundaries.

Though there is considerable variation in the accuracy of various gene prediction programs depending on data sets and the availability and choice of homolog, we believe that a judicious use of these programs in combination can result in highly accurate gene structures for genes with known homologs. There is, however, still considerable progress to be made on predicting alternative spliced structures and genes with no known homologs.

METHODS

Computational Gene Identification Tools

Gene identification tools may be categorized into ab initio tools (those not utilizing sequence similarity and relying on intrinsic gene measures such as coding potential and splice signals), and those based (at least partly) on sequence similarity.

Ab initio Gene Identification Tools

The ab initio gene identification tools use information from both the gene signals in the genomic DNA (such as splice sites, start and stop codons, and promoter elements), and the statistical biases in DNA composition that is characteristic of coding regions. There are a number of such programs (for reveiws, see Fickett 1996; Claverie 1997; Guigó 1997a; Burge and Karlin 1998; Haussler 1998). GENSCAN (Burge and Karlin 1997) is one of the most accurate and widely used programs in this category, and we use it as a representative.

SSBGP Tools

A number of recent programs predict genes by aligning genomic sequences with candidate homologous protein sequences. These programs may include a splice site model, coding potential, and sequence similarity to known proteins to infer gene predictions. We evaluated two of these programs, PROCRUSTES (Gelfand et al. 1996), and GENEWISE (Birney and Durbin 1997) (http://www.sanger.ac.uk/Software/Wise2/).

These programs require as input a candidate homologous protein sequence; therefore, in typical use, a sequence similarity database search with the query genomic sequence is performed a priori and the top hit is used as the candidate (or top hits are used as candidates, in the case of a query sequence encoding multiple genes). The database similarity searches were performed against the nonredundant protein sequence database from NCBI, *nr*, using BLASTX (Altschul et al. 1990; Gish and Sates 1993). BLASTX performs a translation of the query sequence into the six frames, and searches for similarities between each of these translations and the protein sequences in the database.

BLASTX was designed as a similarity-based gene prediction tool, and it is possible to model a gene directly from the database search results. BLASTX, however, does not confine its similarity to exon;, thus the similarity region is not constrained to begin or end on splice sites. Moreover, BLASTX does not explicitly predict genes in genomic sequences, and some postprocessing of its output is required to infer gene predictions from the search results. Indeed, while computational gene finders predict genes, that is pairs of positions (corresponding to exon starts and ends) along the query genomic sequence, database searches only produce lists of sequence database hits along the query sequence. Each hit above a given similarity threshold may be assumed to be a coding exon. For different database entries, however the set of hits may be different. The problem is then to infer a gene model from the set of database hits. A simple solution is to project the hits into a single axis along the genomic sequence, and to assume the union of these projections to be the coding exons.

In total, three strategies based on BLAST were tested:

(1) default — A procedure consisting of projecting the HSPs onto the genomic sequences was used (see Guigó et al. 2000). BLASTX was run with *E = 1e-10 − filter xnu + seg S2 = 60*, and all HSPs with identity <40% were discarded. The choices of S2 and percentage identity were influenced by the need to restrict false matches.

(2) topcomboN — BLASTX was used with default parameters except for *−filter xnu + seg topcomboN = 1*. HSPs with *P* value > 10^{-20} were discarded, and the projections along the query sequence of the remaining HSPs assumed to be the predicted coding exons. WashU–BLAST has a parameter topcomboN that limits all HSPs generated to be in one consistent group. For example, for BLASTX searches, each region of the nucleotide sequence is only aligned to a single region on the protein sequence and the ordering of these HSPs has to be consistent along both the nucleotide and protein sequences. This restricts spurious matches arising from repetitive domains with query sequences, and from low scoring hits in introns and flanking regions.

(3) two-stage — BLASTX was used in a two stage process that first identifies one or more candidate protein sequences in the presence of a low-complexity filter. In the second stage, BLASTX is used to align the candidates individually with the genomic sequence, this time without the filter and with topcomboN = 1. This two pass technique is closer to the strategy used with GENEWISE and PROCRUSTES, where a first BLASTX search pinpoints the protein homolog to be used, and a subsequent GENEWISE uses this protein homolog.

Both GENEWISE and PROCRUSTES were run with mostly standard parameters: GENEWISE v2.1.16b *-both -gff -pretty -para -cdna -genes -quiet* and PROCRUSTES was run in the local mode with *MIN__EXN 20, MIN__IVS*

50, GAP 2, INI__GAP 10, MATRIX pam120.mtx. GENSCAN was run with default parameters.

Benchmark Sets

Two sets of sequences have been used to evaluate the programs discussed above. First, a typical benchmark set made of sequences from the EMBL database release 50 (1997) that included 178 human genomic sequences coding for single complete genes for which both the mRNA and the coding exons are known. The procedure used to extract the sequences is described in Burset and Guigó (1996) and Guigó (1997b). We will refer to this set here as h178. All the genes in this data set are on the forward strand. Other characteristics of h178 are provided in Table 4.

For the reasons discussed in this paper, this does not appear to be a challenging benchmark set for estimating the accuracy of gene identification programs in the larger genomic sequences. Unfortunately, very few large genomic sequences have been studied extensively to produce complete experimental determinations of the exact structure of each gene. To overcome this limitation, we generated a semiartificial set of genomic sequences in which accurate gene annotation can be guaranteed.

In essence, a set of annotated genic sequences are placed randomly in a background of random intergenic DNA. The length of the semiartificial sequence is generated randomly according to a normal distribution. Genomic fragments containing genes and random-sized segments of intergenic sequence are then concatenated until their combined lengths exceed the target. The strands are also chosen at random for each genic subsequence.

Table 4 shows the characteristics of the generated sequences when the method is applied to the sequences in h178 and the intergenic background is generated using a Markov Model of order 5 as described in Guigó and Fickett (1995) assuming an average intergenic G + C content of 38%. The 178 genic sequences were collapsed into 42 SAG sequences. Some of the resulting parameters, such as average G + C content of 40%, a gene every 43 Kb, and a coding density of 2.3% are in agreement with that for the overall human genome. This data set has flaws and is not a perfect representative of the human genome. Some of the ignored characteristics include the isochore organization of the human genome, known and unknown repeats in the intergenic regions, presence of pseudogenes and other evolutionary remnants, genes with huge introns, and tandem gene clusters. Most of the missing properties (pseudogenes, repeats, huge introns) make gene prediction much more difficult. Thus, we expect the accuracy results on Gen178 to still be an overestimate of the true accuracy.

Evaluating Accuracy

The measures of accuracy used here are discussed extensively in Burset and Guigó (1996). We will restate them briefly. Accuracy is measured at three different levels: nucleotide, exon, and gene. At the nucleotide and exon levels, we essentially compute the proportion of actual coding nucleotides/exons that have been predicted correctly–(which we call Sensitivity) and the proportion of predicted coding nucleotides/exons that are actually coding nucleotides/exons (which we call Specificity). To compute these measures at the exon level, we will assume that an exon has been predicted correctly only when both its boundaries have been predicted correctly. To summarize both Sensitivity and Specificity, we compute the Correlation Coefficient at the nucleotide level, and the average of Sensitivity and Specificity at the exon level. At the exon and gene level, we also compute the Missing Exons/Genes (the proportion of actual exons/genes that overlap no predicted exon/gene) and the Wrong Exons/Genes (the proportion of predicted exons/genes that overlap no actual exon/gene).

The measures are computed globally from the total number of prediction successes and failures (at the base and exon level) on all sequences. Accuracy in Table 1 is computed ignoring predictions in the reverse (wrong) strand. The first column in Tables 1 and 2 indicates the number of sequences for which the progams produced predictions.

Data Availability

Both the set of single gene sequences and the set of semiartificially generated genomic sequences will be available from http://www1.imim.es/databases/gpecal2000/.

ACKNOWLEDGMENTS

We thank Randall F. Smith, Ewan Birney, Chris Burge, and Warren Gish, and the anonymous referees (one in particular for pointing out the topcomboN feature in WU-BLAST) for useful comments. This work was partially supported by a grant from Plan Nacional de I + D, BIO98-0443-C02-01, and from the Ministerio de Educación y Ciencia (Spain) to R.G. M.B. is supported by a Formación de Personal Investigador fellowship, FP95-38817943, from the Ministerio de Educación y Ciencia (Spain), J.F.A. is supported by a predoctoral fellowship, 99/9345, from the Instituto de Salud Carlos III (Spain).

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be

Table 4. Characteristics of the Benchmark Sequence Sets

			Sequence length			Genes (average)				CDS (average)		
Set	No.	G + C	average	min	max	no.	length	density		no. exons	length	density
h178	178	50%	7169	622	86640	1	3657	53%	7169	5.1	968	21%
Gen178	42	40%	177160	70037	282097	4.1	15136	8.6%	43000	21	4007	2.3%

The columns Genes (average) and CDS (average) provide values averaged over all the sequences (178 in h178 and 42 in Gen178). Gene density provides the percentage of nucleotides that occur in genic regions (exons, introns, and UTRs), and the number of kilobases per gene. CDS no. exons is the average number of coding exons per sequence, and CDS density is the percentage of nucleotides that occur in coding regions.

hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

Abril, J.F. and Guigó, R. 2000. gff2ps: A tool for visualizing genomic annotations. *Bioinformatics* in press.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D. 1990. Basic local alignment search tool. *J. Mol. Biol.* **215:** 403–410.

Birney, E. and Durbin, R. 1997. Dynamite: A flexible code generating language for dynamic programming methods used in sequence comparison. *Ismb* **5:** 56–64.

Burge, C.B. and Karlin, S. 1997. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* **268:** 78–94.

———. 1998. Finding the genes in genomic DNA. *Curr. Opin. Struc. Biol.* **8:** 346–354.

Burset, M. and Guigó, R. 1996. Evaluation of gene structure prediction programs. *Genomics* **34:** 353–357.

Church, D.M., Stotler, C.J., Rutter, J.L., Murrell, J.R., Trofatter, J.A., and Buckler, A.J. 1994. Isolation of genes from complex sources of mammalian genomic DNA using exon amplification. *Nat. Genet.* **6:** 98–105.

Claverie, J.-M. 1997. Computational methods for the identification of genes in vertebrate genomic sequences. *Hum. Mol. Genet.* **6:** 1735–1744.

Couch, F.J., Rommens, J.M., Neuhausen, S.L., Couch, E.J., Rommens, J.M., Neuhausen, S.L., Belanger, C., Dumont, M., Abel, K., Bell, R., Berry, S., Bogden, R., Cannon-Albright, L. 1996. Generation of an integrated transcription map of the BRCA2 region on chromosome 13q12-q13. *Genomics* **36:** 86–99.

Fickett, J.W. 1996. Finding genes by computer: the state of the art. *Trends Genet.* **12:** 316–320.

Gelfand, M.S., Mironov, A.A., and Pevzner, P.A. 1996. Gene recognition via spliced alignment. *PNAS* **93:** 9061–9066.

Gish, W. and States, D. 1993. Identification of protein coding regions by database similarity search. *Nat. Genet.* **3:** 266–272.

Guigó, R. 1997a. Computational gene identification. *J. Mol. Med.* **75:** 389–393.

———. 1997b. Computational gene identification: An open problem. *Comput. Chem.* **21:** 215–222.

Guigó, R. and Fickett, J.W. 1995. Distinctive sequence features in protein coding, genic non-coding, and inter-genic human DNA. *J. Mol. Biol.* **253:** 51–60.

Guigó, R., Burset, M., Agarwal, P., Abril, J.F., Smith, R.F., and Fickett, J.W. 2000. Sequence similarity based gene prediction. In *Genomics and proteomics: Functional and computational aspects* (ed. S. Suhai), pp. 95–105. Kluwer Academic / Plenum Publishing, New York, NY.

Haussler, D. 1998. Computational genefinding. In *Trends Biochem. Sci., supplementary guide to bioinformatics*, pp. 12–15.

Hochgeschwender, U. 1992. Toward a transcriptional map of the human genome. *Trends Genet.* **8:** 41–44.

Huang, X., Adams, M.D., Zhou, H., and Kerlavage, A.R. 1997. A tool for analyzing and annotating genomic sequences. *Genomics* **46:** 37–45.

Krogh, A. 1997. Two methods for improving performance of an HMM and their application for gene finding. *ISMB* **5:** 179–186.

Kulp, D., Haussler, D., Reese, M.G., and Eeckman, F.H. 1996. A generalized hidden markov model for the recognition of human genes in DNA. In *Intelligent systems for molecular biology* (eds. D.J. States, P. Agarwal, T. Gaasterland, L. Hunter, and R. Smith), pp. 134–142. AAAI Press, Menlo Park, CA.

Kulp, D., Haussler, D., Reese, M.G., and Eeckman, F.H. 1997. Integrating database homology in a probabilistic gene structure mode. In *Biocomputing: Proceedings of the 1997 Pacific Symposium* (eds. R.B. Altman, A.K. Dunke, L. Hunter, and T.E. Klein), pp. 232–244. World Scientific, New York, NY.

Mott, R. 1997. EST__GENOME: A program to align spliced DNA sequences to unspliced genomic DNA. *Comput. Appl. Biosci.* **13:** 477–478.

Reese, M.G., Hartzell, G., Harris, N.L., Ohler, U., Abril, J.F., and Lewis, S.E. 2000. Genome annotation assessment in *Drosophila melanogaster*. *Genome Res.* **10:** 483–501.

The *C. elegans* Sequencing Consortium. 1998. Genome sequence of the nematode *C. elegans*: A platform for investigating biology. *Science* **282:** 2012–2018.

Received October 12, 1999; accepted in revised form August 11, 2000.

3.3.4 Reese *et al*, *Genome Research*, 10(4):483–501, 2000

PubMed Accession:

http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&list_uids=10779488&dopt=Abstract

Journal Abstract:

http://www.genome.org/cgi/content/abstract/10/4/483

Supplementary Materials:

http://www.fruitfly.org/GASP1/
http://genome.imim.es/datasets/Dro_me/
http://genome.imim.es/software/gfftools/GFF2PS-ADHposter.html

Companion Poster:

See Figure 3.7 and the following URLs:
http://www.genome.org/content/vol10/issue4/images/data/483/DC1/GR10n4_poster.zip
http://genome.imim.es/references/genome_maps/2000_GenomeResearch_v10_i4_p483_poster_GASP.ps.gz

Letter

Genome Annotation Assessment in *Drosophila melanogaster*

Martin G. Reese,[1,4] George Hartzell,[1] Nomi L. Harris,[1] Uwe Ohler,[1,2] Josep F. Abril,[3] and Suzanna E. Lewis[1]

[1]Berkeley Drosophila Genome Project, Department of Molecular and Cell Biology, University of California, Berkeley, California 94720-3200 USA; [2]Chair for Pattern Recognition, University of Erlangen–Nuremberg, D-91058 Erlangen, Germany;[3]Institut Municipal d'Investigació Médica—Universitat Pompeu Fabra, Department of Medical Informatics (IMIM—UPF), 08003 Barcelona, Spain

Computational methods for automated genome annotation are critical to our community's ability to make full use of the large volume of genomic sequence being generated and released. To explore the accuracy of these automated feature prediction tools in the genomes of higher organisms, we evaluated their performance on a large, well-characterized sequence contig from the *Adh* region of *Drosophila melanogaster*. This experiment, known as the Genome Annotation Assessment Project (GASP), was launched in May 1999. Twelve groups, applying state-of-the-art tools, contributed predictions for features including gene structure, protein homologies, promoter sites, and repeat elements. We evaluated these predictions using two standards, one based on previously unreleased high-quality full-length cDNA sequences and a second based on the set of annotations generated as part of an in-depth study of the region by a group of *Drosophila* experts. Although these standard sets only approximate the unknown distribution of features in this region, we believe that when taken in context the results of an evaluation based on them are meaningful. The results were presented as a tutorial at the conference on Intelligent Systems in Molecular Biology (ISMB-99) in August 1999. Over 95% of the coding nucleotides in the region were correctly identified by the majority of the gene finders, and the correct intron/exon structures were predicted for >40% of the genes. Homology-based annotation techniques recognized and associated functions with almost half of the genes in the region; the remainder were only identified by the ab initio techniques. This experiment also presents the first assessment of promoter prediction techniques for a significant number of genes in a large contiguous region. We discovered that the promoter predictors' high false-positive rates make their predictions difficult to use. Integrating gene finding and cDNA/EST alignments with promoter predictions decreases the number of false-positive classifications but discovers less than one-third of the promoters in the region. We believe that by establishing standards for evaluating genomic annotations and by assessing the performance of existing automated genome annotation tools, this experiment establishes a baseline that contributes to the value of ongoing large-scale annotation projects and should guide further research in genome informatics.

Genome annotation is a rapidly evolving field in genomics made possible by the large-scale generation of genomic sequences and driven predominantly by computational tools. The goal of the annotation process is to assign as much information as possible to the raw sequence of complete genomes with an emphasis on the location and structure of the genes. This can be accomplished by ab initio gene finding, by identifying homologies to known genes from other organisms, by the alignment of full-length or partial mRNA sequences to the genomic DNA, or through combinations of such methods. Related techniques can also be used to identify other features, such as the location of regulatory elements or repetitive sequence elements. The ultimate goal of genome annotation, the functional classification of all the identified genes, currently depends on discovering homologies to genes with known functions.

We are interested in an objective assessment of the state of the art in automated tools and techniques for annotating complete genomes. The Genome Annotation Assessment Project (GASP) was organized to formulate guidelines and accuracy standards for evaluating computational tools and to encourage the development of new models and the improvement of existing approaches through a careful assessment and comparison of the predictions made by current state-of-the-art programs.

The GASP experiment, the first of its kind, was similar in many ways to the CASP (Critical Assessment of techniques for protein Structure Prediction) contests for protein structure prediction (Dunbrack et al. 1997;

[4]Corresponding author.
E-MAIL mgreese@lbl.gov; FAX (510) 486-6798.

Levitt 1997; Moult et al. 1997, 1999; Sippl et al. 1999; Zemla et al. 1999), described at http://predictioncenter.llnl.gov. However, unlike the CASP contest, GASP was promoted as a collaboration to evaluate various techniques for genome annotation.

The GASP experiment consisted of the following stages: (1) Training data for the *Adh* region, including 2.9 Mb of *Drosophila melanogaster* genomic sequence, was collected by the organizers and provided to the participants; (2) a set of standards was developed to evaluate submissions while the participating groups produced and submitted their annotations for the region; and (3) the participating groups' predictions were compared with the standards, a team of independent assessors evaluated the results of the comparison, and the results were presented as a tutorial at ISMB-99(Reese et al. 1999).

Participants were given the finished sequence for the *Adh* region and some related training data, but they did not have access to the full-length cDNA sequences that were sequenced for the paper by Ashburner et al. (1999b) that describes the *Adh* region in depth. The experiment was widely announced and open to any participants. Submitters were allowed to use any available technologies and were encouraged to disclose their methods. Because we were fortunate to attract a large group of participants who provided a wide variety of annotations, we believe that our evaluation addresses the state of art in genome annotation.

Twelve groups participated in GASP, submitting annotations in one or more of six categories: ab initio gene finding, promoter recognition, EST/cDNA alignment, protein similarity, repetitive sequence identification, and gene function. Table 1 lists each participating group, the names of the programs or systems it used, and which of the six classes of annotations it submitted (see enclosed poster in this issue for a graphic overview of all the groups' results). Additional papers in this issue are written by the participants themselves and describe their methods and results in detail.

It should be noted that the lack of a standard that is absolutely correct makes evaluating predictions problematic. The expert annotations described by the *Drosophila* experts (Ashburner et al. 1999b) are our best available resource, but their accuracy will certainly improve as more data becomes available. At best, the data we had in hand is representative of the true situation, and our conclusions would be unchanged by using a more complete data set. At worst, there is a bias in the available data that makes our conclusions significantly misleading. We believe that the data is not unreasonable and that conclusions based on it are correct enough to be valuable as the basis for discussion and future development. We do not believe that the values for the various statistics introduced below are precisely what they would be using the extra information, and we emphasize that they should always be considered in the context of this particular annotated data set [for a further detailed discussion of evaluating these predictions, see Birney and Durbin (2000)].

In the next section we describe the target genomic sequence and the auxiliary data, including a critical discussion of our standard sets. Methods gives a short

Table 1. Participating Groups and Associated Annotation Categories

	Program name	Gene finding	Promoter recognition	EST/c DNA alignment	Protein similarity	Repeat	Gene function
Mural et al. Oakridge, US	`GRAIL`	X		X			X
Parra et al. Barcelona, ES	`GeneID`	X					
Krogh Copenhagen, DK	`HMMGene`	X					
Henikoff et al. Seattle, US	`BLOCKS`				X		X
Solovyev et al. Sanger, UK	`FGenes`	X					
Gaasterland et al. Rockefeller, US	`MAGPIE`	X	X	X		X	X
Benson et al. Mount Sinai, US	`TRF`					X	
Werner et al. Munich, GER	`CoreInspector`		X				
Ohler et al. Nuremberg, GER	`MCPromoter`		X				
Birney Sanger, UK	`GeneWise`				X		X
Reese et al. Berkeley/Santa Cruz, US	`Genie`	X	X				

description of existing annotation methods that complements other papers in this issue, including a review article of existing gene-finding methods by Stormo (2000) and papers describing the methods used by the individual participants. Results assesses the individual annotation methods and the Conclusions discusses what the experiment revealed about issues involved in annotating complete genomes. An article by Ashburner (2000) provides a biological perspective on the experiment.

Data: The Benchmark Sequence: The *Adh* Region in *D. melanogaster*

The selection of a genomic target region for assessing the accuracy of computational genome annotation methods was a difficult task for several reasons: The genomic region had to be large enough, the organism had to be well studied, and enough auxiliary data had to be available to have a good experimentally verified "correct answer," but the data should be anonymous so that a blind test would be possible. The *Adh* region of the *D. melanogaster* genome met these criteria. *D. melanogaster* is one of the most important model organisms, and although the *Adh* region had been extensively studied, the best gene annotations and cDNAs for the region were not published until after the conclusion of the GASP experiment. The 2.9 Mb *Adh* contig was large enough to be challenging, contained genes with a variety of sizes and structures, and included regions of high and low gene density. It was not a completely blind test, however, because several cDNA and genomic sequences for known genes in the region were available prior to the experiment.

Genomic DNA Sequence

The contiguous genomic sequence of the *Adh* region in the *D. melanogaster* genome spans nearly 3 Mb and has been sequenced from a series of overlapping P1 and BAC clones as a part of the Berkeley Drosophila Genome Project (BDGP; Rubin et al. 1999) and the European Drosophila Genome Project (EDGP; Ashburner et al. 1999c). This sequence is believed to be of very high quality with an estimated error rate of <1 in 10,000 bases, based on PHRAP quality scores. A detailed analysis of this region can be accessed through the BDGP Web site (http://www.fruitfly.org/publications/Adh.html) as well as in Ashburner et al. (1999b).

Curated Training Sequences

We provided several *D. melanogaster*-specific data sets to the GASP participants. This enabled participants to tune their tools for *Drosophila* and facilitated a comparison of the various approaches that was unbiased by organism-specific factors. The following curated sequence sets, extracted from Flybase and EMBL (provided by the EDGP at Cambridge and provided by the BDGP, were made available and can be found at http://www.fruitfly.org/GASP/data/data.html): (1) A set of complete coding sequences (start to stop codon), excluding transposable elements, pseudogenes, noncoding RNAs, and mitochondrial and viral sequences (2122 entries); (2) nonredundant set of repetitive sequences, not including transposable elements (96 entries); (3) transposon sequences, containing only the longest sequence of each transposon family and excluding defective transposable elements (44 entries); (4) genomic DNA data from 275 multi- and 141 single-exon nonredundant genes together with their start and stop codons and splice sites, taken from GenBank version 109; (5) a set of 256 unrelated promoter regions, taken from the Eukaryotic Promoter Database (EPD; Cavin Périer et al. 1999, 2000) and a collection made by Arkhipova (1995); and (6) an uncurated set of cDNA and EST sequences from work in progress at the BDGP. Five of the 12 participating groups reported making use of these data sets.

Resources for Assessing Predictions: The "Correct" Answer

In a comparative study, the gold standard used to evaluate solutions is the most important factor in determining the usefulness of the study's results. For the results to be meaningful, the standard must be appropriate and correct in the eyes of the study's audience. Because our goal was to evaluate tools that predict genes and gene structure in complex eukaryotic organisms, we drew our standard from a complex eukaryotic model organism, choosing to work with a 2.9-Mb sequence contig from the *Adh* region of *D. melanogaster*. Comparing predicted annotations in such a region is only consequential if the standard is believed to be correct, if that correctness has been established by techniques that are independent of the approaches being studied, and if the predictors had no prior knowledge of the standard. Ideally, it would contain the correct structure of all the genes in the region without any extraneous annotations. Unfortunately, such a set is impossible to obtain because the underlying biology is incompletely understood. We built a two-part approximation to the perfect data set, taking advantage of data from the BDGP cDNA sequencing project (http://www.fruitfly.org/EST) and a *Drosophila* community effort to build a set of curated annotations for this region (Ashburner et al. 1999b). Our first component, known as the std1 data set, used high-quality sequence from a set of 80 full-length cDNA clones from the *Adh* region to provide a standard with annotations that are very likely to be correct but certainly are not exhaustive. The second component, known as the std3 data set, was built from the annotations being developed for Ashburner et al.(1999b) to give a standard with more complete coverage of the region, although with less confidence about the accuracy and independence of the annotations. We believe that this two-part approximation allows us to draw useful conclusions about the ability to accurately predict gene structure in complex eukaryotic organisms even though the absolutely perfect data set does not exist.

Eukaryotic transcript annotations have complex structures based on the composition of fundamental features such as the TATA box and other transcription factor binding sites, the transcription start site (TSS), the start codon, 5′ and 3′ splice site boundaries, the stop codon, the polyadenylation signal, exon start and end positions, and coding exon start and end positions. Our gene prediction evaluations focused on annotations that are specific to the coding region, from the start codon through the various intron–exon boundaries to the stop codon, and on promoter annotations. Although other types of features are also biologically interesting, we were unable to devise reliable methods for evaluating their predictions. Whenever possible, we relied on unambiguous biological evidence for our evaluations; when that was not available, we combined several types of evidence curated by domain experts.

Our goal for our first standard set, called std1, was to build a set of annotations that we believed were very likely to be correct in their fine details (e.g., exact locations for splice sites), even if we were unable to include every gene in the region. We based std1 on alignments of 80 high-quality, full-length cDNA sequences from this region with the high-quality genomic sequence for the contig. The cDNA sequences are the product of a large cDNA sequencing project at the BDGP and had not been submitted to GenBank at the time of the experiment. Working from five cDNA libraries, the longest clone for each unique transcript was selected and sequenced to a high-quality level. Starting with these cDNA sequences, we generated alignments to the genomic sequence using sim4 (Florea et al. 1998) and filtered them on several criteria. Of the 80 candidate cDNA sequences, 3 were paralogs of genes in the *Adh* region and 19 appeared to be cloning artifacts (unspliced RNA or multiple inserts into the cloning vector), leaving us with alignments for 58 cDNA clones. These alignments were further filtered based on splice site quality. We required that all of the proposed splice sites include a simple "GT"/"AG" core for the 5′ and 3′ splice sites, respectively, and that they scored highly (5′ splice sites ≥ 0.35 threshold, which gives a 98% true positive rate, and 3′ splice sites ≥ 0.25, which gives a 92% true positive rate) using a neural network splice site predictor trained on *D. melanogaster* data (Reese et al. 1997). This process left us with 43 sequences from the *Adh* region for which we had structures confirmed by alignments of high-quality cDNA sequence data with high-quality genomic data and by the fit of their splice sites to a *Drosophila* splice site model. Of these 43 sequences, 7 had a single coding exon and 36 had multiple coding exons. We added start codon and stop codon annotations to these structures from the corresponding records in the std3 data set.

After the experiment, we recently discovered four inconsistent genes in the std1 data set. For two genes (*DS07721.1*, *DS003192.4*), the cDNA clones (CK02594, CK01083, respectively) are likely to be untranscribed genomic DNA that was inappropriately included in the cDNA library. Two other genes from std3 (*DS00797.5* and *wb*) were incorrectly reported in std1 as three partial all incomplete EST alignments (cDNA clones: CK01017, LD33192, and CK02229). In keeping with std1's goal of highly reliable annotations, all four sequences have been removed from the std1 data set that is currently available on the GASP web site. The results reported here use the larger, less reliable, data set as presented at the ISMB-99 tutorial.

The complete set of the original 80 aligned high-quality, full-length cDNA sequences was named std2. This set was never used in the evaluation process because it did not add any further compelling information or conclusions because of the unreliable alignments.

Our goal for the second, used standard set, called std3, was to build the most complete set of annotations possible while maintaining some confidence about their correctness. Ashburner et al. (1999b) compiled an exhaustive and carefully curated set of annotations for this region of the *Drosophila* genome based on information from a number of sources, included BLASTN, BLASTP (Altschul et al. 1990), and PFAM alignments (Sonnhammer et al. 1997, 1998; Bateman et al. 2000), high scoring GENSCAN (Burge and Karlin 1997) and Genefinder (P. Green, unpubl.) predictions, ORFFinder results (E. Friese, unpubl.), full-length cDNA clone alignments (including those used in std1), and alignments with full-length genes from GenBank. This set included 222 gene structures: 39 with a single coding exon and 183 with multiple coding exons. Of these 222 gene structures, 182 are similar to a homologous protein in another organism or have a *Drosophila* EST hit. For these structures, the intron–exon boundaries were verified by partial cDNA/EST alignments using sim4 (Florea et al. 1998), homologies were discovered using BLASTX, TBLASTX, and PFAM alignments, and gene structure was verified using a version of GENSCAN trained for finding human genes. Of the 54 remaining genes, 14 had EST or homology evidence but were not predicted by GENSCAN or Genefinder, and 40 were based entirely on strong GENSCAN and Genefinder predictions. All of this evidence was evaluated and edited by experienced *Drosophila* biolo gists, resulting in a protein coding gene data set that exhaustively covers the region with a high degree of confidence and represents their view of what should or should not be considered an annotated gene. Their gene data set excluded the 17 found transposable elements [6 LINE-like elements (*G*, *F*, *Doc*, and *jockey*) and 11 retrotransposons with long terminal repeats (LTRs; *copia*, *roo*, *297*, *blood*, *mdg1*-like, and *yoyo*)], which almost all contain long ORFs. Some of these ORFs code for known and some others for, so far, unknown protein sequences.

Both of these data sets have shortcomings. As mentioned above, std1 only includes a subset of the genes in the region. It also includes a pair of transcripts that represent alternatively spliced products of a single gene. Although this is not incorrect, it confounds our scoring process. Because the cDNA alignments do not provide any evidence for the location of the start and stop codons, we based those annotations in std1 on information from the std3 set. Many of the gene structures in std3 are based on GENSCAN and Genefinder predictions without other supporting evidence, so it is possible that the fine details are incorrect, that the entries are not entirely independent of the techniques used by the predictors in the experiment, and that the set overestimates the number of genes in the region.

See Birney and Durbin (2000) and Henikoff and Henikoff (2000) for further discussion of the difficulties of evaluating these predictions especially in the protein homology annotation category, in which, by training, these programs will recognize protein-like sequences such as the ORFs in transposable elements as genes. They and others (see other GASP publications in this issue) have raised the issues of annotation oversights, transposons, and pseudogenes. In cases where GASP submissions suggest a missed annotation, this information has been passed onto biologists for further research, including screening cDNA libraries. We believe that it would have been biased to retroactively change the scoring scheme used at the GASP experiment based solely on missed annotations discovered by the participant's submissions. See Discussion for an example of an annotation that may be missing in the standard data sets. In the std3 data set we based our standard for what is or is not a *Drosophila* gene on the expert annotations provided by Ashburner et al. (1999b). It is clear that both transposons and pseudogenes are genuine features of the genome and that gene-finding technologies might recognize them. Because they were not included as coding genes in the expert annotations, we decided against including them in the standard set.

Building a set for the evaluation of transcription start site or, more generally, for promoter recognition proved to be even more difficult. For the genes in the *Adh* region almost no experimentally confirmed annotation for the transcription start site exists. As the 5′ UTR regions in *Drosophila* can extend up to several kilobases, we could not simply use the region directly upstream of the start codon. To obtain the best possible approximation, we took the 5′ ends of annotations from Ashburner et al. (1999b) where the upstream region relied on experimental evidence (the 5′ ends of full-length cDNAs) and for which the alignment of the cDNA to the genomic sequence included a good ORF. The resulting set contained 92 genes of the 222 annotations in the std3 set (Ashburner et al. 1999b). This number is larger than the number of cDNAs used for the construction of the std1 set described above because we included cDNAs that were already publicly available. The 5′ UTR of these 96 genes has an average length of 1860 bp, a minimum length of 0 bp (when the start codon was annotated at the beginning, due to the lack of any further cDNA alignment information; this is very likely to be only a partial 5′ UTR and therefore an annotation error), and a maximum length of 36,392 bp.

Data Exchange Format

One of the challenges of a gene annotation study is finding a common format in which to express the various groups' predictions. The format must be simple enough that all of the groups involved can adapt their software to use it and still be rich enough to express the various annotations.

We found that the General Feature Format (GFF) (formerly known as the Gene Feature Finding format) was an excellent fit to our needs. The GFF format is an extension of a simple *name*, *start*, *end* record that includes some additional information about the sequence being annotated: the source of the feature; the type of feature; the location of the feature in the sequence; and a score, strand, and frame for the feature. It has an optional ninth field that can be used to group multiple predictions into single annotations. More information can be found at the GFF web site: http://www.sanger.ac.uk/Software/formats/GFF/. Our evaluation tools used a GFF parser for the PERL programming language that is also available at the GFF web site.

We found that it was necessary to specify a standard set of feature names within the GFF format, for instance, declaring that submitters should describe coding exons with the feature name CDS. We produced a small set of example files (accessible from the GASP web site) that we distributed to the submitters and were pleased with how easily we were able to work with their results.

METHODS

Genome annotation is an ongoing effort to assign functional features to locations on the genomic DNA sequence. Traditionally, most of these annotations record information about an organism's genes, including protein-coding regions, RNA genes, promoters, and other gene regulatory elements, as well as gene function. In addition to these gene features, the following general genome structure features are also commonly annotated: repetitive elements and general A, C, G, T content measures (e.g., isochores).

Genome Annotation Classes

Although the GASP experiment invited and encouraged any class of annotations, most submissions were for gene-related features, emphasizing ab initio gene predictions and promoter predictions. In addition, two groups submitted functional protein domain annotations, and two groups submitted repeat element annotations. In the sections that follow, we categorize and discuss the submitted predictions.

Gene Finding

Protein coding region identification is a major focus of computational biology. A separate article in this issue (Stormo 2000) discusses and compares current methods, whereas an early paper by Fickett and Tung (1992) and a more recent review of gene identification systems by Burge and Karlin (1998) give excellent overviews of the field. Table 2 lists the six groups that predicted protein-coding regions with the corresponding program names. It also categorizes the submissions based on the types of information used to build the model for predictions. Although all groups used statistical information for their models—predominantly coding bias, coding preference, and consensus sequences for start codon, splice sites, and stop codons—only two groups used protein similarity information or promoter information to predict gene structure. More than half of the groups incorporated sequence information from cDNA sequences. In general, state-of-the-art gene prediction systems use complex models that integrate multiple gene features into a unified model.

Promoter Prediction

The complicated nature of the transcription initiation process makes computational promoter recognition a hard problem. We define promoter prediction as the identification of TSSs of protein coding genes that are transcribed by eukaryotic RNA polymerase II. A detailed description of the structure of promoter regions and existing promoter prediction systems is beyond the scope of this paper. Fickett and Hatzigeorgiou (1997) provide an excellent review of the field.

We can broadly identify three different approaches to promoter prediction, with at least one GASP submission in each category. The first class consists of "search by signal" programs, which identify single binding sites of proteins involved in transcription initiation or combinations of sites to improve the specificity. The program CoreInspector by Werner's group (M. Scherf, A. Klingenhoff, and T. Werner, in prep.) belongs to this category and searches for co-occurrences of two common binding sites within the core promoter (the core promoter usually denotes the region where the direct contact between the transcription machinery, the holoenzyme of the transcription complex, and the DNA takes place). The second class is often termed "search by content," as programs within this group do not rely on specific signals but take the more general approach of identifying the promoter region as a whole, frequently based on statistical measures. Sometimes the promoter is split into several regions to obtain more accurate statistics. The MCPromoter program (Ohler et al. 1999) is a member of this second group. In comparison with the signal-based group, the content-based systems usually are more sensitive but less specific. The third class can be described as "promoter prediction through gene finding." Simply using the start of a gene prediction as a putative TSS can be very successful if the 5′ UTR region is not too large. This approach can be improved by including similarity to EST sequences and/or a promoter module in the statistical systems used for gene prediction. The TSS predictions submitted by the participants of the MAGPIE and the Genie groups belong to this last class.

The notorious difficulty of the problem itself is exacerbated by the limited amount of existing reliably annotated training material. The experimental mapping of a TSS is a laborious process and is therefore not routinely carried out, even if the gene itself is studied extensively. So, both training the models and evaluating the results is a difficult task, and the conclusions we draw from the results must be considered with much caution.

Repeat Finders

Detecting repeated elements plays a very important role in modeling the three-dimensional structure of a DNA molecule, specifically, the packing of the DNA in the cell nucleus. It is believed that the packing of the DNA around the nucleosome is correlated with the global sequence structure produced predominantly by repetitive elements. Repeats also play a major role in evolution (for review, see Jurka 1998). Two groups, Gary Benson [tandem repeats finder v. 2.02 (TRF; Benson 1999)] and the MAGPIE team using two programs Calypso (D. Field, unpubl.) and REPuter (Kurtz and Schleiermacher 1999) submitted repetitive sequence annotations. TRF (Benson 1999) locates approximate tandem repeats (i.e., two or more contiguous, approximate copies of a pattern of nucleotides) where the pattern size is unspecified but falls within the range from 1 to 500 bases. The Calypso program (D. Field, unpubl.) is an evolutionary genomics program. Its primary function is to find repetitive regions in DNA and protein sequences that have higher than average mutation rates. The REPuter program (Kurtz and Schleiermacher 1999) determines repeats of a fixed preselected length in complete genomes.

Table 2. Gene-Finding Submissions

	Program name	Statistics	Promoter	EST/cDNA alignment	Protein similarity
Mural et al. (Oakridge, US)	GRAIL	X		X	
Guigó et al. (Barcelona, ES)	GeneID	X			
Krogh (Copenhagen, DK)	HMMGene	X		X	X
Solovyev et al. (Sanger, UK)	FGenes	X			
Gaasterland et al. (Rockefeller, US)	MAGPIE	X	X	X	
Reese et al. (Berkeley/Santa Cruz, US)	Genie	X	X	X	X

Protein Homology Annotation

Homologies to gene sequences from other organisms can often be used to identify protein-coding regions in anonymous genomic sequence. In addition to the location, it is often possible to infer the function of the predicted gene based on the function of the homologous gene in the other organism or of a known structural and functional protein element in the gene. Whereas the tools in the gene prediction category and the EST/cDNA alignment category are usually intended to determine the exact structure of a gene, the protein homology-based tools are usually optimized to find conserved parts of the sequence without worrying about the exact gene structure. Traditionally, this area of genome annotations has been dominated by the suite of local alignment search tools of `BLAST` (Altschul et al. 1990) and more global search tools such as `FASTA` (Pearson and Lipman 1988). Recent reviews in this area include Agarwal and States (1998), Marcotte et al. (1999), and Pearson (1995).

In the GASP experiment, two groups specializing in functional protein domain or motif identification in genomic DNA submitted annotations. The Henikoff group found hits to the `BLOCKS+` database (http://blocks.fhcrc.org), a database consisting of conserved protein motifs (Henikoff and Henikoff 1994; Henikoff et al. 1999a). The second group in this category submitted results from the `GeneWise` program (Birney 1999). This program searches genomic DNA against a comprehensive hidden Markov model (HMM)-based library (`PFAM`; Sonnhammer et al. 1997, 1998; Bateman et al. 2000) of protein domains. Both programs look for conserved regions by searching translated DNA against a representation of multiple aligned sequences. Whereas in `BLOCKS+` the multiple protein alignments consist of sets of ungapped regions, the `GeneWise` program searches against a gapped alignment. Both methods will turn up distantly related sequences.

EST/cDNA Alignment

Computational predictions of gene location and structure go hand in hand with EST/cDNA sequencing and alignment techniques for building transcript annotations in genomic sequence. Either can be used as a discovery tool, with the other held in reserve for verification. A researcher can verify the existence and structure of predicted genes by sequencing the corresponding mRNA molecules and aligning their sequences to the original genomic sequence. Alternatively, one can start with an EST or cDNA sequence and build an alignment to the genomic sequence that has been guided and/or verified by tools from the gene prediction arsenal, for example, using likely splice site locations and checking for long ORFs and potential frame shifts.

There are many tools for aligning sequences. Although they have generally been specialized for aligning sequences that are evolutionarily related, some are designed for niche applications such as recognizing overlaps among sequencing runs. Aligning EST/cDNA sequences to the original genomic sequence also presents a unique set of tradeoffs and issues. In some cases (interspecies EST/genomic alignments), these tools must model evolutionary changes in the sequence. Sometimes (e.g., for low-quality EST sequences), they need to model errors in the sequence generated by the sequencing process. For multiexon genes, they need to model the intron regions as cost-free gaps tied to a model for recognizing splice sites. Several tools have been developed for this task: Mott (1997) and Birney and Durbin (1997) describe dynamic programming approaches that include models of splice sites and intron gaps. Florea et al. (1998) describe `sim4`, a heuristic tool that performs as well as the dynamic programming approaches and is efficient enough to support searching of large databases of genomic sequence.

Using cDNA clones and their sequences to build transcript annotations requires a variety of operations. The tools discussed above align the cDNA sequences to the genomic sequence, but steps must be taken to filter out clones that are merely paralogs of genes in the sequence and to recognize and handle various laboratory artifacts. If the clones represent short ESTs, then a likely annotation can be built by assembling a consistent model from their individual alignments. Longer ESTs or cDNAs might generate several similar alignments, and an automated tool must be able to select the most biologically meaningful variant. Although there are some gene prediction tools that can use information about homologies to known genes or ESTs, and most large-scale sequencing centers have some automated sanity checking for their database search results, there are not any tools that automate the production of transcript annotations from cDNA sequences.

Gene Function

Gene function predictions are the most difficult annotations to produce and to evaluate. Current technologies use similarity to proteins (or protein domains) with known function to predict functional domains in genomic sequence. Although some tools use simple sequence alignments, more powerful tools have developed significantly more sensitive models.

It quickly became apparent that a consistent and correct assessment of function predictions as part of the GASP experiment was not possible because of the incomplete understanding of the protein products encoded by the 222 genes in the *Adh* region.

Evaluating Gene Predictions

An ideal gene prediction tool would produce annotations that were exactly correct and entirely complete. The fact that no existing tool has these characteristics reflects our incomplete understanding of the underlying biology as well as the difficulty to build adequate gene models in a computer. Although no tool is perfect, each tool has particular strengths and weaknesses, and any performance evaluation should be in the context of an intended use. For example, researchers who are interested in identifying gene-rich regions of a genome for sequencing would be happy with a tool that successfully recognizes a gene's approximate location, even if it incorrectly described splice site boundaries. On the other hand, someone trying to predict protein structures is more interested in getting a gene's structure exactly right than in a tool's ability to predict every gene in the genome.

When assessing the accuracy of predictions, each prediction falls into one of four categories. A true-positive (TP) prediction is one that correctly predicts the presence of a feature. A false-positive (FP) prediction incorrectly predicts the presence of a feature. A true-negative (TN) prediction is correct in not predicting the presence of a feature when it isn't there. A false-negative (FN) prediction fails to predict the existence of a feature that actually exists. The sensitivity (Sn) of a tool is defined as TP / (TP + FN) and can be thought of as a measure of how successful the tool is at finding things that are really there. The specificity (Sp) of a tool is defined as TP / (TP + FP) and can be thought of as a measure of how careful a tool is about not predicting things that aren't really there. Burset and Guigó (1996) also use a correlation coefficient and an average correlation coefficient. We chose not to use these measures because they depend on predictors' TN information, and we recognize that our evaluation sets were constructed in such a way that the TN information is not trustworthy. These Sn and Sp metrics are used for evaluating the submissions in the gene-finding, promoter recognition, and gene identification using protein homology categories. In the gene finding category, they are used for all three levels: base level, exon level, and gene level. In the protein homology category, they are used for base level and gene level only.

In one of the first reviews of gene prediction accuracy, Fickett and Tung (1992) developed a method that measured predictors' ability to correctly recognize coding regions in genomic sequence. They used their method to compare published techniques and concluded that in-frame hexamer counts were the most accurate measure of a region's coding potential. Burset and Guigó (1996) recognized that there are a wide variety of uses for gene predictions and developed measures—including base level, exon level, and gene level Sp and Sn—that describe a predictor's suitability for a particular task.

Base Level

The base level score measures whether a predictor is able to correctly label a base in the genomic sequence as being part of some gene. It rewards predictors that get the broad sweeps of a gene correct, even if they don't get the details such as the splice site boundaries entirely correct. It penalizes predictors that miss a significant portion of the coding sequence, even if they get the details correct for the genes they do predict. We used the Sn and Sp measures defined above as the measures of success in this category.

Exon Level

Exon level scores measure whether a predictor is able to identify exons and correctly recognize their boundaries. Being off by a single base at either end of the exon makes the prediction incorrect. Because we only considered coding exons in our assessment, the first exon is bracketed by the start codon and a 5′ splice site, the last exon is bracketed by a 3′ splice site and the stop codon, and the interior exons are bracketed by a pair of splice sites. As measures of success in this category, we used two statistics in addition to Sn and Sp. The missed exon (ME) score is a measure of how frequently a predictor completely failed to identify an exon (no prediction overlap at all), whereas the wrong exon (WE) score is a measure of how frequently a predictor identifies an exon that has no overlap with any exon in the standard sets. The ME score is the percentage of exons in the standard set for which there were no overlapping exons in the predicted set. Similarly, the WE score is the percentage of exons in the predicted set for which there were no overlapping exons in the standard set.

Gene Level

Gene level Sn and Sp measure whether a predictor is able to correctly identify and assemble all of a gene's exons. For a prediction to be counted as a TP, all of the coding exons must be identified, every intron–exon boundary must be exactly correct, and all of the exons must be included in the proper gene. This is a very strict measure that addresses a tool's ability to perfectly identify a gene. In addition to the Sn and Sp measures based on absolute accuracy, we used the missed genes (MG) score as a measure of how frequently a predictor completely missed a gene (a standard gene is considered missed if none of its exons are overlapped by a predicted coding gene) and the wrong genes (WG) score as a measure of how frequently a predictor incorrectly identified a gene (a prediction is considered wrong if none of its exons are overlapped by a gene from the standard set).

Split and Joined Genes

The exon level scores discussed above measure how well a predictor recognizes exons and gets their boundaries exactly correct. The gene level scores measure how well a predictor can recognize exons and assemble them into complete genes. Neither of these scores directly measures a predictor's tendency to incorrectly assemble a set of predicted exons into more or fewer genes than it should. We developed two new measures, split genes (SG) and joined genes (JG), which describe how frequently a predictor incorrectly splits a gene's exons into multiple genes and how frequently a predictor incorrectly assembles multiple genes' exons into a single gene. Because the coverage of the std1 data set is so incomplete, we have only included SG and JG scores from the comparison with std3. A gene from the standard set is considered split if it overlaps more than one predicted gene. Similarly, a predicted gene is considered joined if it overlaps more than one gene in the standard set. The SG measure is defined as the sum of the number of predicted genes that overlap each standard gene divided by the number of standard genes that were split. Similarly, the JG measure is the sum of the number of standard genes that overlap each predicted gene divided by the number of predicted genes that were joined. A score of 1 is perfect and means that all of the genes from one set overlap exactly one gene from the other set.

Application of These Measures to Correct Answer Data Sets std1/std3

We built the std1 data set in such a way that we believe it is correct in the details of the genes that it describes, though we know that it only includes a small portion of the genes in the region. The std3 data set, on the other hand, is as complete as was possible but does not have rigorous independent evidence for all of its annotations. For the std1 data set, we believe that the TP count (it was predicted, and it exists in the standard) and FN count (it was not predicted, but it does exist in the standard) are reliable because of the confidence that we have in the correctness of the predictions in the set. On the other hand, we do not believe that the TN count (it was not predicted, and it is not in the standard set) and FP count (it was predicted, but is not in the standard set) are reliable because they both assume that the standard correctly describes the absence of a feature and we know that there are genes missing from std1. It follows that we believe that Sn is meaningful for std1 because it only depends on TP and FN but that we are less confident about the Sp score because it depends on TP and FP. A similar logic applies to the std3 data set, where our confidence in the set's completeness but not its fine details suggests that the TP and FP scores are usable but that the TN and FN scores are not. This means that for std3, we believe that the Sp measure can be used to describe a predictor's performance but that Sn is likely to be misleading.

Evaluation of Promoter Predictions

We adopted the measures proposed by Fickett and Hatzigeorgiou (1997). They evaluated the success of promoter predictions by giving the percentage of correctly identified TSSs versus the FP rate. A TSS is regarded as identified if a program makes one or more predictions within a certain "likely" region around the annotated site. The FP rate is defined as the number of predictions within the "unlikely" regions outside the likely regions divided by the total number of bases contained in the unlikely set. As our annotation of the TSS is only preliminary and not experimentally confirmed, we chose a rather large region of 500 bases upstream and 50 bases downstream of the annotated TSS as the likely region. The upstream region is always taken as the likely region, even if it overlaps with a neighboring gene annotation on the same strand. The unlikely region for each gene then consists of the rest of the gene annotation, from base 51 downstream of the TSS to the end of the final exon.

Visualization of the Annotations

Generating "good" annotations generally requires integrating multiple sources of information, such as the results of various sequence analysis tools plus supporting biological information. Visualization tools that display sequence annotations in a browsable graphical framework make this process much more efficient. In this experiment we found that visualization tools are essential to evaluate the genome annotation submissions. When annotations are displayed visually, overall trends become apparent, for example, gene-rich versus gene-poor regions, genes that were predicted by most participants versus those that were predicted by few. Additionally, as we discuss below, a visualization tool that is capable of displaying annotations at multiple levels of detail provides a way to examine individual predictions in detail.

Building genome annotation visualization tools is a daunting task. Many such tools have been developed, starting with ACeDB (Eeckman and Durbin 1995; Stein and Thierry-Mieg 1998). We were fortunate in that the BDGP has built a flexible suite of genome visualization tools (Helt et al. 1999) that could be extended to display the GASP submissions. We adapted the BDGP's annotated clone display and editing tool, `CloneCurator` (Harris et al. 1999), which is based on a genomic visualization toolkit (Helt et al. 1999), to read the annotation submissions in `GFF` format and display each team's predictions in a unique color and location.

`CloneCurator` (see Fig. 1) displays features on a sequence as colored rectangles. Features on the forward strand appear above the axis, whereas those on the reverse strand appear below the axis. The display can be zoomed and scrolled to view areas of interest in more detail. A configuration file identifies the feature types that are to be displayed and assigns colors and offsets to each one. For example, the std1 and std3 exons appear in yellow and orange close to the central axis.

RESULTS

The results of an experiment such as GASP are only meaningful if enough groups participate. We were fortunate to have 12 diverse groups involved, and we were very grateful for the speed with which they were able to submit their predictions. We believe that these 12 groups provide a fair representation of the state of the art in annotation system technology. We collected submissions by electronic mail and evaluated them using the std1 and std3 data sets as described above. Before releasing our results at the Intelligent Systems in Molecular Biology conference in August 1999 in Heidelberg, Germany, we assembled a team of independent assessors (Ashburner et al. 1999a) to review our techniques and conclusions. As discussed in the introduction, the accuracy of the various measures discussed below depends heavily on how well our standard sets capture the true set of features in the region. These values should only be considered in the context of the standard data sets.

A detailed description of the results and the evaluation techniques we used can be accessed through the GASP homepage at http://www.fruitfly.org/GASP/.

Gene Finding

Table 3 summarizes the performance of the gene-finding tools using the measures defined above. Three groups submitted multiple submissions. The first group, `Fgenes1`, `Fgenes2`, and `Fgenes3`, submitted three predictions at varying stringency (for details, see Salamov and Solovyev 2000). For the `GeneID` program, two submitted versions are presented, version 1 (`GeneID v1`) being the original submission and version 2 (`GeneID v2`) being a newer submission from a corrected version of the original program (for details, see Parra et al. 2000). The third group with multiple submissions used three versions of the `Genie` program: the first a pure statistical approach (`Genie`), the second including EST alignment information (`GenieEST`), and the third using protein homology information (`GenieESTHOM`) (for details, see Reese et al. 2000). For all other groups from Table 2, only one submission was evaluated. The following sections discuss the base level, exon level, and gene level performance of these submissions.

Base Level Results

Several gene prediction tools had a Sn of >0.95 at the base level. This suggests that current technology is able to correctly identify >95% of the *D. melanogaster* proteome. A few tools demonstrated a specificity of >0.90 at the base level, only infrequently labeling a noncoding base as coding. Generally, the tools have a higher Sn than Sp. Two programs, `Fgenes2` and `GeneID`, were designed to be conservative about their predictions and do not follow this trend.

Exon Level Results

There was a great deal of variability in the exon level scores. Several tools had Sn scores ~0.75, correctly identifying both exon boundaries ~75% of the time. Their Sps were generally much lower (the highest was 0.68), probably a reflection of the strict definition of exon level scores both splice sites had to be predicted correctly and possible inaccuracies in the std3 data set. The low ME scores (several <0.05) combined with the fairly high Sn suggest that several tools were successful at identifying exons but had trouble finding the correct exon boundaries. Programs that incorporate EST alignment information, such as `GenieEST` and `HMMGene`, had sensitivity scores that were up to 10% bet-

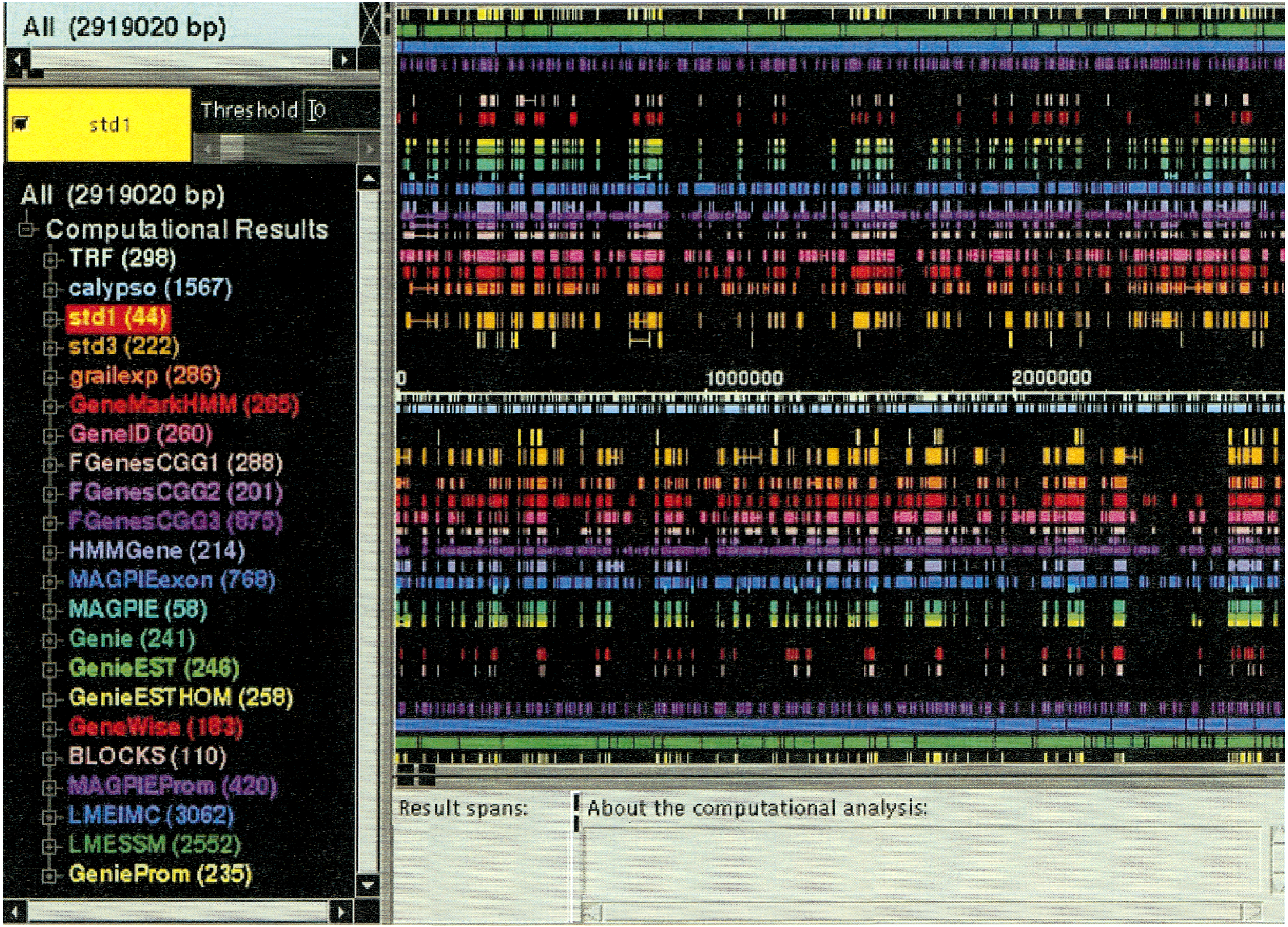


Figure 1 (*See facing page for legend.*)

ter than the other tools. The high WE scores suggest either that the tools are overpredicting or that there are genes that are missing even from std3.

Gene Level Results

All of the predictors had considerable difficulty correctly assembling complete genes. The best tools were able to achieve Sns between 0.33 and 0.44, meaning that they are incorrect over half of the time. This value seems to be very similar in *Drosophila* and human sequences, based on a recent analysis of the *BRCA2* region in human (T.J. Hubbard, pers. comm.). Even on the more complete std3 data set, the programs tended to incorrectly predict many genes. The very low MG score (as low as 4.6%) is reassuring because it suggests that several tools are able to recognize a gene, even if they have difficulty figuring out the exact details of its structure. Comparing the WG and MG measures suggests that existing tools tend to predict genes that do not exist more often than they miss genes that do exist. Because it is almost certain that there are real genes that are missing from both standard sets, this conclusion must be viewed with some skepticism. Although there were several tools with good SG or JG scores, none of them performed well in both categories.

Promoter Prediction

Table 4 shows the performance of the promoter prediction systems, grouped by approach: search-by-signal, search-by-region, and gene prediction programs.

Gene-finding programs that include a prediction of the TSS obtained the best results. The number of false predictions made by the region-based programs is very high (giving them a low Sp), and because the signal-specific programs only identify one promoter, their Sn is very low. The high Sp of the gene finders is obviously due to the context information: All promoter predictions within gene predictions are ruled out in advance, and the location of the possible start codon provides the system with a good initial guess of where to look for a promoter. The `MAGPIE` system also uses EST alignments to obtain information on 5′ UTRs, which mirrors the way the std sets were constructed: Roughly one-third of the putative TSS assignments rely on cDNAs that were publicly available in GenBank. A closer look at the results reveals that the region-based programs have a Sn that is comparable with the gene finders and the signal based program had only a single FP, showing that both types of tools can be used for different applications.

Our data set, and the evaluation based on it, relies on the assumption that the 5′ ends of the full-length cDNAs are reasonably close to the TSS. This makes it very hard to draw strong conclusions from the pre-

Figure 1 (*See facing page.*) Screen shot from the `CloneCurator` program (Harris et al. 1999), featuring the genome annotations of all 12 groups for the 2.9-Mb *Adh* region. The main panel shows the computational annotations on the forward (above axis) and reverse sequence strands (below axis). Genes located on the *top* half of each map are transcribed from distal to proximal (with respect to the telomere of chromosome are 2L); those on the *bottom* are transcribed from proximal to distal. Right below the axis are the two repeat finding results displayed, followed by reference sets from Ashburner et al. (1999b; std1 and std3), followed by the 12 submissions of gene-finding programs, followed by the two protein homology programs, and eventually, farthest away from the axis, the four promoter recognition programs. (*Left*) The color-coded legend for the program and the number of predictions made by the programs.

Program identifier	Color	Reference
`TRF`	seafoam	Benson (1999)
`Calypso`	lightblue	D. Field (unpubl.)
`std1`	yellow	unpublished conservative alignment of cDNAs
`std3`	orange	Ashburner et al. (1999b)
`Grailexp`	red-orange	Uberbacher and Mural (1991)
`GeneMarkHMM`	red	Besemer and Borodovsky (1999)
`GeneID`	hotpink	Guigó et al. (1992)
`FGenesCGG1`	pink	Solovyev et al. (1995)
`FGenesCGG2`	magenta	Solovyev et al. (1995)
`FGenesCGG3`	purple	Solovyev et al. (1995)
`HMMGene`	cornflower	Krogh (1997)
`MAGPIEexon`	blue	Gaasterland and Sensen (1996)
`MAGPIE`	turquoise	Gaasterland and Sensen (1996)
`Genie`	seagreen	Reese et al. (1997)
`GenieEST`	green	Kupl et al. (1997)
`GenieESTHOM`	chartreuse	Kulp et al. (1997)
`GeneWise`	red	Birney (1999)
`BLOCKS`	pink	Henikoff et al. (1999b)
`MAGPIEProm`	purple	T. Gaasterland, (unpubl.)
`LMEIMC`	blue	Ohler et al. (1999)
`LMESSM`	dark green	Ohler et al. (2000)
`GeniePROM`	chartreuse	Reese (2000)

Table 3. **Evaluation of Gene-Finding Systems**

		FGenes 1	FGenes 2	FGenes 3	GeneID v1	GeneID v2	Genie	Genie EST	Genie ESTHOM	HMMGene	MAGPIE exon	GRAIL
Base level	Sn std1	0.89	0.49	0.93	0.48	0.86	0.96	0.97	0.97	0.97	0.96	0.81
	Sp std3	0.77	0.86	0.60	0.84	0.83	0.92	0.91	0.83	0.91	0.63	0.86
Exon level	Sn std1	0.65	0.44	0.75	0.27	0.58	0.70	0.77	0.79	0.68	0.63	0.42
	Sp std3	0.49	0.68	0.24	0.29	0.34	0.57	0.55	0.52	0.53	0.41	0.41
	ME (%) std1	10.5	45.5	5.6	54.4	21.1	8.1	4.8	3.2	4.8	12.1	24.3
	WE (%) std3	31.6	17.2	53.3	47.9	47.4	17.4	20.1	22.8	20.2	50.2	28.7
Gene level	Sn std1	0.30	0.09	0.37	0.02	0.26	0.40	0.44	0.44	0.35	0.33	0.14
	Sp std3	0.27	0.18	0.10	0.05	0.10	0.29	0.28	0.26	0.30	0.21	0.12
	MG (%) std1	9.3	34.8	9.3	44.1	13.9	4.6	4.6	4.6	6.9	4.6	16.2
	WG (%) std3	24.3	24.8	52.3	22.2	30.5	10.7	13.0	15.5	14.9	55.0	23.7
	SG	1.10	1.10	2.11	1.06	1.06	1.17	1.15	1.16	1.04	1.22	1.23
	JG	1.06	1.09	1.08	1.62	1.11	1.08	1.09	1.09	1.12	1.06	1.08

The evaluation is divided into three categories: base level, exon level, and gene level. The different statistical features reported are Sn, Sp, ME, WE, MG, WG, SG, and JG. std1 and std3 indicate against which standard set the statistics are reported.

sented results. Even the most sensitive systems could identify only roughly one third of the start sites. This could of course be caused by the fact that the existing annotation is only an approximation and some of the true TSSs may be located further upstream. It also hints at the diversity of promoter regions that mirrors the possibilities for gene regulation and at the existing bias toward housekeeping genes in the current data sets used for the training of the models.

Gene Identification Using Protein Homology

Gene-finding evaluation statistics, such as those described above, can be used to summarize the ability of a program to identify complete and accurate gene structures in genomic DNA. In Table 5 we have applied the same evaluation statistics to the homology-based search programs GeneWise and BLOCKS+. Because these programs are not optimized to deal with exact exon boundary assignments, Table 5 only shows the performance for the base level and the MG and WG.

The very low Sns at the base level are not surprising, because the programs identify only conserved protein motifs or particular domains and make no effort to predict complete genes. Sp, which should be high given that only conserved protein motifs are scored, was lower than expected. Detailed studies of these pre-

Table 4. **Evaluation of Promoter Prediction Systems**

System name	Sensitivity	Rate of false-positive predictions in region[a] (853,180 bases)	Rate of predictions in region[b] (2,570,232 bases)
CoreInspector	1 (1%)	1/853,180	1/514,046
MCPromoter v1.1	26 (28.2%)	1/2,633	1/2,537
MCPromoter v2.0	31 (33.6%)	1/2,437	1/2,323
GeniePROM	25 (27.1%)	1/14,710	1/28,879
GenieESTPROM	30 (32.6%)	1/16,729	1/29,542
MAGPIE	33 (35.8%)	1/14,968	1/16,370

We show the Sn for identified TSSs in comparison with the FP rate for non-TSS regions and general gene regions: [a]the unlikely region defined as the rest of the gene starting 51 bases downstream from its annotated TSS; [b]the general gene region, spanning from half the distance to the previous and next annotated genes including the annotated TSS (taken from the std3 annotation).

Table 5. Evaluation of Similarity Searching

		BLOCKS	GeneWise	MAGPIE cDNA	MAGPIE EST	Grail Similarity
Base level	Sn std1	0.04	0.12	0.02	0.31	0.31
	Sp std3	0.80	0.82	0.55	0.32	0.81
Gene level	MG (%) std1	62.7	69.7	95.3	27.9	41.8
	WG (%) std3	12.9	14.1	0.0	44.3	7.4

Base and gene level statistics are shown. The base level is described using Sn and Sp, and the statistics for the gene level are given as MG and WG.

dictions (see Birney and Durbin 2000; Henikoff and Henikoff 2000) show that most of the FP predictions were hits to transposable elements or to possible genes that are missing in the standard sets. Both programs use a database of protein domains or conserved protein motifs. Both databases are large and are believed to contain at least 50% of the existing protein domains. The high number of MG, 62.7% for BLOCKS and 69.7% for GeneWise, means that these programs will miss a significant number of *Drosophila* genes when used to search genomic DNA directly. The WG scores of 12.9% BLOCKS and 14.1% for GeneWise are lower than the gene finding programs discussed in the previous section.

Gene Identification Using EST/cDNA Alignments

It is believed that some cDNA information exists for approximately half of the genes in the *D. melanogaster* genome. This cDNA database (available as the EST data set at the GASP web site) was used as a basis for the cDNA/EST alignment category. The Sn of 31% for MAGPIEEST and GrailSimilarity (Table 5) implies that the coding portion of the available EST data currently covers one-third of the genome's coding sequence. The low Sp is very surprising and suggests that the EST/cDNA alignment problem is not a trivial one. The only program that tried to align complete cDNAs to genomic DNA, MAGPIEcDNA, could find complete cDNAs for only 2.4% of the genes. EST alignments also resulted in high numbers of missed genes, suggesting that the EST libraries are biased toward highly expressed genes. The high WG scores suggest that some genes are missing even from std3.

Selected Gene Annotations

The summary statistics discussed above only provide a global view of the predicting programs' characteristics. A much better understanding of how the various approaches behave can be obtained by looking at individual gene annotations. Such a detailed examination can also help identify issues that are not addressed by current systems.

In the following paragraphs we will discuss a few interesting examples. Figure 1 shows the color codes of the participating groups that are used throughout this section. Genes located on the top of each map are transcribed from distal to proximal (with respect to the telomere of chromosome arm 2L); those on the bottom are transcribed from proximal to distal. std1 and std3 are the expert annotations described in Ashburner et al.(1999b). Just below the axis, you can see the annotations for the two repeat finding programs. These have no sequence orientation and are therefore only shown on one side. Farther away from the axis, after std1 and std3, we grouped all of the ab initio gene-finding programs together. Next to the gene finders are the homology-based annotations. On the bottom and the top of the figure we show the three promoter annotations, but for clarity we did not include these annotations in the subsequent figures. (On the front page and in the legend of Fig. 1, you can see the full set of annotations of all programs, which are also accessible from the GASP web site.)

Our first example is a "busy" region with 12 complete genes and 1 partial gene in a stretch of only 40 kb (Fig. 2A). This region is located at the 3′ end of the *Adh* region from base 2,735,000 to base 2,775,000. Genes exist on both strands, and it is striking that in this region the genes tend to alternate between the forward and the reverse strands. We selected this region for its gene density and because it has characteristics that are typical of the complete *Adh* region. Figure 2A vividly demonstrates that all of the gene-finding programs' predictions are highly correlated with the annotated genes from std1/std3. In the past, gene finders had often mistakenly predicted a gene on the noncoding strand opposite of a real gene, leading to FP predictions known as "shadow exons." Figure 2A makes it clear that gene finders have overcome this problem, because there are almost no shadow exon predictions for any of the genes in std3. Another characteristic, captured in the high base level sensitivity and the low missing genes statistics, is that every gene in the std3 set was predicted by at least a few groups and that most of these predictions agree with each other. Except for the second and third genes [*DS02740.5*, *I(2)35Fb*] on the forward strand (2,740,000–2,745,000), which seem to

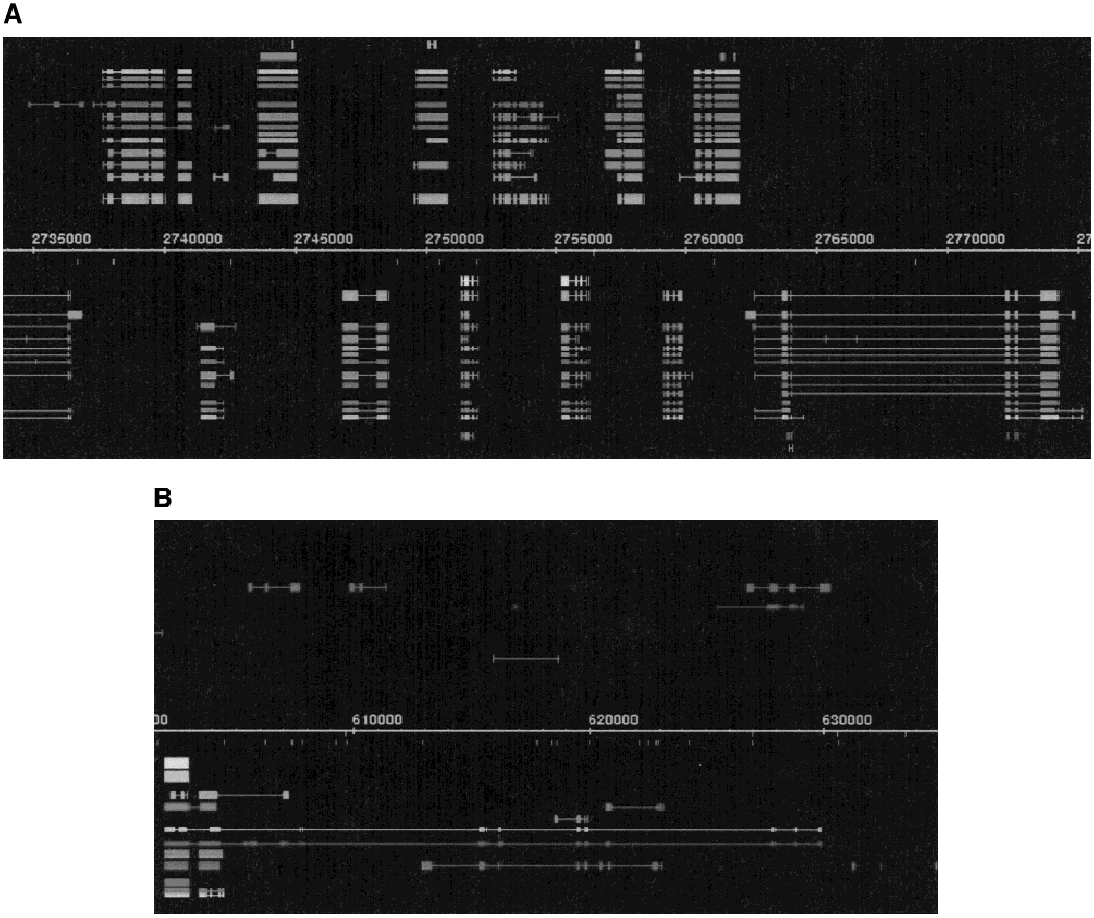


Figure 2 (*A*) Annotations for the following known genes described in Ashburner et al. (1999b) are shown for the region from 2,735,000 to 2,775,000 (from the *left* to the *right* of the map): *crp* (partial, reverse (r)), *DS02740.4* (forward (f)), *DS02740.5* (f), *l(2)35Fb* (f), *heix* (r), *DS02740.8* (f), *DS02740.9* (r), *DS02740.10* (f), *anon-35Fa* (r), *Sed5* (f), *cni* (r), *fzy* (f), *cact* (r). (*B*) Annotations for the following known gene described in Ashburner et al. (1999b) are shown for the region from 600,000 to 635,000 (*left* to *right*): *DS01759.1* (r).

be single exon genes, all of the genes in this region are multiexon genes with between two and eight exons. The exon size varies widely. There are genes that consist of only two large exons, some that consist of a mix of large and small exons, and some that are made up exclusively of many small exons. The distribution seems to be almost random. Except for the long final intron in the last gene on the reverse strand (*cact*), the region consists exclusively of short introns.

Predictions on the reverse strand indicate a possible gene from base 2,741,000 to base 2,745,000. Most of the gene finders agree on this prediction, but neither std1 nor std3 describes a gene at this location. This could be a real gene that was missed by the expert annotation pathway described in Ashburner et al (1999b). Neither BLOCKS+ nor GeneWise found any homologies in this region, but we can see from the table in the previous section that many real genes do not have any homology annotations. Interestingly, this is the only area in the region where two gene finders predicted a possible gene that likely consists of shadow exons.

The fifth gene on the forward strand (*DS02740.10*, bases 2,752,500–2,755,000) shows that long genes with multiple exons are much harder to predict than single exon genes or genes with only a few exons. In this region splitting and joining genes does not seem to be a problem. Repeats occur sparsely and mostly in noncoding regions, predominantly in introns.

In contrast to the busy region in Figure 2A, Figure

2B highlights a region of almost equal size in which only one gene (*DS01759.1*) is present in both std1 and std3. There are very few FP predictions by any group, but there is one case where the "false" predictions by different programs are located at very similar positions (on the reverse strand near base 620,000). This suggests a real gene that is missing from both standard sets.

Figure 3, A–D, depicts selected genes that illustrate some interesting challenges in gene finding. Figure 3A shows the *Adh* and the *Adhr* genes that occur as gene duplicates. The encoded proteins have a sequence identity of 33%. The positions of the two introns interrupting the coding regions are conserved and give additional evidence to tandem duplication. Both genes are under the control of the same regulatory promoter, the *Adhr* gene does not have a TSS of its own, and its transcript is always found as part of an *Adh–Adhr* dicistronic mRNA. Gene duplications occur very frequently in the *Drosophila* genome—estimates show that at least 20% of all genes occur in gene family duplications. In an additional twist, *Adh* and *Adhr* are located within an intron of another gene, *outspread* (*osp*), that is found on the opposite strand (for details, see Fig. 3B). The *Adh* gene is correctly predicted by most of the programs, although one erroneously predicts an additional first exon. Most of the programs also predict the structure of *Adhr* correctly; one program misses the initial exon and shortens the second exon. Both *Adh* and *Adhr* show hits to the protein motifs in BLOCKS+ as well as alignments to a PFAM protein domain family through GeneWise. Both genes hit two different PFAM families, and the order of these two domains is conserved in the gene structure.

Figure 3B highlights the *osp* gene region. This is an example of a gene with exceptionally long (>20 kb) introns, making it hard for any gene finder to predict the entire structure correctly. In addition, there are a number of smaller genes [including the *Adh* and *Adhr* genes discussed above, *DS09219.1* (r.) and *DS07721.1* (f.)] within the introns of *osp*. No current gene finder includes overlapping gene structures in its model; as a consequence, none of the GASP gene finders were able to predict the *osp* structure without disruption. This is clearly a shortcoming of the programs because genes containing other genes are often observed in *Drosophila* (Ashburner et al. 1999b report 17 cases for the *Adh* region). However, it should be noted that most of the gene finders predict the 3′ end of *osp* correctly and therefore get most of the coding region right. The region that includes the 5′ end of *osp* shows a lot of gene prediction activity, but there is no consistency among the predictions. One program (FGenesCCG3) does correctly predict the *DS09219.1* gene.

Figure 3C shows the entire gene structure of the *Ca-α1D* gene. This gene is the most complex gene in the *Adh* region, with >30 exons. This is a very good example for studying gene splitting. Several predictors break the gene up into several genes, but some groups make surprisingly close predictions. This shows the complex structure that genes can exhibit and that extent to which this complexity has been captured in the state-of-the-art prediction models. It is interesting to note that most of the larger exons are predicted, whereas the shorter exons are missed. Such a large complex gene is a good candidate for alternative splicing, which can ultimately be detected only by extensive cDNA sequencing.

Figure 3D shows the triple duplication of the *idgf* gene (*idgf1*, *idgf2*, and *idgf3*) on the forward strand. Two programs mistakenly join the first two genes into a single gene; all the others correctly predict all three genes.

DISCUSSION

The goal of the GASP experiment was to review and assess the state of the art in genome annotation tools. We believe that the noncompetitive framework and the community's enthusiastic participation helped us achieve that goal. By providing all of the participants with an unprecedented set of *D. melanogaster* training data and using unreleased information about the region as our gold standard, we were able to establish the level playing field that made it possible to compare the performance of the various techniques. The large size of the *Adh* contig and the diversity of its gene structures provided us with an opportunity to compare the capabilities of the annotation tools in a setting that models the genome-wide annotations currently being attempted. However, the lack of a completely correct standard set means that our results should only be considered in the context of the std1 and std3 data sets.

Assessing the Results

The most difficult part of the assessment was developing a benchmark for the predicted annotations. By dividing the predictions into different classes and developing class-specific metrics that were based on the best available standards, we feel that we were able to make a meaningful evaluation of the submissions. Although most of the information that was used to evaluate the submissions was unreleased, some cDNA sequences from the region were in the public databases. As sequencing projects move forward, it will become increasingly difficult for future experiments to find similarly unexplored regions. This makes it very different from the CASP protein structure prediction contests, which can use the three-dimensional structure of a novel target protein that is unknown to the predictors.

As discussed in the introduction, the lack of an absolutely correct standard against which to evaluate the various predictions is a troubling issue. Although we believe that the standard sets sufficiently represent

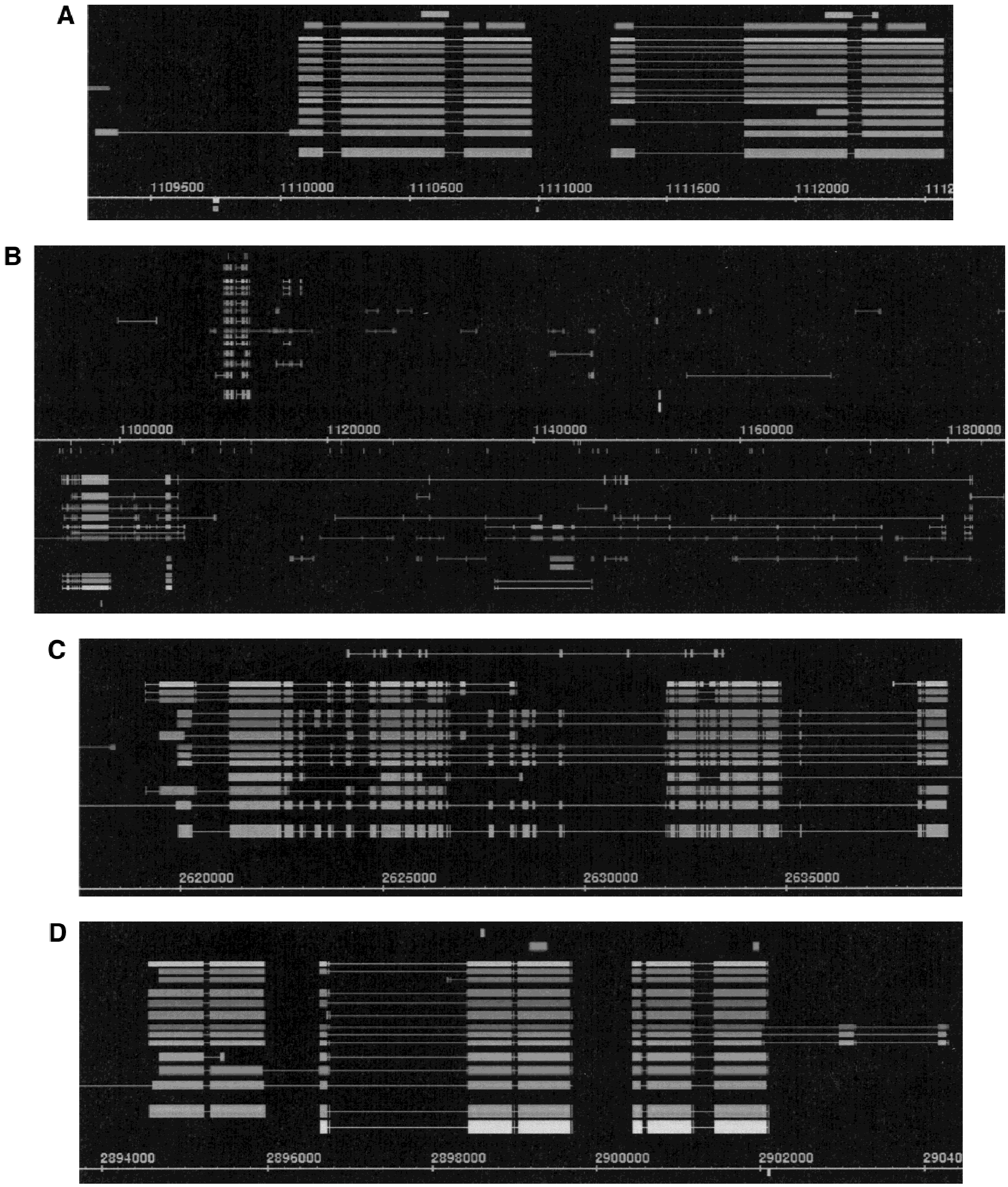


Figure 3 (*A*) Annotations for the following known genes described in Ashburner et al. (1999b) are shown for the region from 1,109,500 to 1,112,500 (forward strand only) (*left* to *right*): *Adh, Adhr*. (*B*) Annotations for the following known genes described in Ashburner et al. (1999b) are shown for the region from 1,090,000 to 1,180,000 (*left* to *right*): *osp* (r), *Adh* (f), *Adhr* (f), *DS09219.1* (r), *DS07721.1* (f). (*C*) Annotations for the following known gene described in Ashburner et al. (1999b) are shown for the region from 2,617,500 to 2,640,000 (forward strand only) (*left* to *right*): *Ca-α1D*. (*D*) Annotations for the following known genes described in Ashburner et al. (1999b) are shown for the region from 2,894,000 to 2,904,000 (forward strand only) (*left* to *right*): *idgf1, idgf2, idgf3*.

the true nature of the region and that conclusions based on them are interesting, it must be remembered that the various results can only be evaluated in the context of these incomplete data sets. This also makes GASP more difficult and less clear cut than CASP, where the three-dimensional protein structure is experimentally solved at least to some degree of resolution.

It should also be noted that the gene-finding tools with the highest Sp have a great deal in common with GENSCAN, the gene prediction tool used in the development of the std3 data set. This suggests that std3's origins might have led to a bias favoring GENSCAN-like predictors. Because std1 was exclusively created using full-length cDNA alignments, this set might be biased towards highly expressed genes, because the cDNA libraries were not normalized.

Progress in Genome-Wide Annotation

The rapid release of completed genomes, including the imminent release of the *D. melanogaster* and human genomes, has driven significant developments in genome annotation and gene-finding tools. Problems that have plagued gene-finding programs, such as predicting shadow exons, restricting predictions to a single strand, recognizing repeats, and accurately identifying splice sites, have been overcome by the current state of the art. In this section, we discuss some of the remaining issues in genome annotation that the GASP experiment highlighted.

Successful gene prediction programs use complex models that integrate information from statistical features that are driven by the three-dimensional protein–DNA/RNA interactions. They make integrated predictions on both strands and have been tuned to predict all the genes in gene-rich regions and avoid overpredicting genes in gene-poor regions (Fig. 2A,B). Although most of the programs identify almost all the existing genes (as evidenced by the Sn and MG statistics), there is significant variation in their ability to accurately predict precise gene structures (see the Sp statistics, particularly at the exon level). If any global performance conclusion can be drawn, it is that the probabilistic gene finders (mostly HMM based) seem to be more reliable. The integration of EST/cDNA sequence information into the ab initio gene finders [see HMMGene, GenieEST, and GRAIL (Fig. 2A,B and Fig. 3A–D)] significantly improves gene predictions, particularly the recognition of intron–exon boundaries. Some groups submitted multiple annotations of the *Adh* region using programs that were tuned for different tasks. The suite of Fgenes programs shows very nicely the results of such a three-part submission. The first Fgenes submission (Fgenes1) is a version adjusted to weight Sn and Sp equally. The second submission (Fgenes2) is very conservative and only annotates high-scoring genes. This results in a high Sp but a low Sn. The third submission (Fgenes3) tries to maximize Sn and to avoid missing any genes, at the cost of a loss in Sp. These differently tuned variants may be useful for different types of tasks.

A comparison (data not shown) to a gene-finding system that was trained on human data showed that it did not perform as well as the programs that were trained on *Drosophila* data.

None of the gene predictors screened for transposable elements, which have a protein-like structure. As described in Ashburner et al. (1999b), the *Adh* region has 17 transposable element sequences. Eliminating transposons from the predictions or adding them to the standard sets would have reduced the FP counts, raising the Sp and lowering the WE and WG scores. Although this accounts for a portion of the high FP scores, we believe that there may also be additional genes in this region not annotated in std3. Future biological experiments (Rubin 2000) to identify and sequence the predicted genes that were not included in std3 should improve the completeness and accuracy of the final annotations.

There were fewer submissions of homology-based annotations than those by ab initio gene finders, and their results were significantly affected by their FP rates. A significant portion of those FPs were matches to transposable elements, some appear to be matches to pseudogenes, and others are likely to be real, but as yet unannotated, genes. The homology-based approaches seem to be the most promising techniques for inferring functions for newly predicted genes.

Even using EST/cDNA alignments to predict gene structures is not as simple as expected. Paralogs, low sequence quality of mRNAs, and the difficulty of cloning infrequently expressed mRNAs make this method of gene finding more complex than believed, and it is difficult to guarantee completeness with this method. Normalized cDNA libraries and other more sophisticated technologies to purify genes with low expression levels, along with improved alignment and annotation technologies, should improve predictions based on EST/cDNA alignments.

Lessons for the Future

To fully assess the submitted annotations, the correct answer must be improved. Only extensive full-length cDNA sequencing can accomplish this. A possible approach would be to design primers from predicted exons and/or genes in the genomic sequence and then use hybridization technologies to fish out the corresponding cDNA from cDNA libraries. For promoter predictions, another way to improve the correct answer is to make genome-to-genome alignments with the DNA of related species (e.g., *Caenorhabditis briggsae* vs. *Caenorhabditis elegans*; *D. melanogaster* vs. *D. virilis*). More detailed guidelines, including how to handle ambiguous features such as pseudogenes and transposons, will make the results of future experiments even more useful.

A successful system to identify all genes in a genome should consist of a combination of ab initio gene finding, EST/cDNA alignments, protein homology methods, promoter recognition, and repeat finding. All of the various technologies have advantages and disadvantages, and an automated method for integrating their predictions seems ideal.

Beyond the identification of gene structure is the determination of gene functions. Most of the existing prototypes of such systems are based on sequence homologies. Although this is a good starting point, it is definitely not sufficient. The state of the art for predicting function in protein sequences uses the protein's three-dimensional structure, but the difficulty of accurately predicting three-dimensional structure from primary sequences makes applying these techniques on complete genomes problematic. The new field of structural genomics will hopefully give more answers in these areas.

Another approach to function classification is the analysis of gene expression data. Improvements in TSS annotations, along with correlation in expression profiles, should be very helpful in identifying regulatory regions.

Conclusions

The GASP experiment succeeded in providing an objective assessment of current approaches to gene prediction. The main conclusions from this experiment are that current methods of gene predictions are tremendously improved and that they are very useful for genome scale annotations but that high-quality annotations also depend on a solid understanding of the organism in question (e.g., recognizing and handling transposons).

Experiments like GASP are essential for the continued progress of automated annotation methods. They provide benchmarks with which new technologies can be evaluated and selected.

The predictions collected in GASP showed that for most of the genes, overlapping predictions from different programs existed. Whether or not a combination of overlapping predictions would do better than the best performing individual program was not explicitly tested in this experiment. For such a test, additional experiments such as cDNA library screening and subsequent full-length cDNA sequencing in this selected *Adh* test bed region would be necessary. These experiments are currently under way, and it would be interesting to perform a second GASP experiment when more cDNAs have been sequenced.

We believe that existing automated annotation methods are scalable and that the ultimate test will occur when the complete sequence of the *D. melanogaster* genome becomes available. This experiment will set standards for the accuracy of genome-wide annotation and improve the credibility of the annotations done in other regions of the genome.

URLs

Gene Finding

HMMGene, http://www.cbs.dtu.dk/services/HMMGene/; GRAIL, http://compbio/ornl.gov/droso; Fgenes, http://genomic/sanger.ac.uk/gf/gf.shtml; GeneID, http://www1/imim.es/~rguigo/AnnotationExperiment/index.html; Genie, http://www.neomorphic. com/genie.

Promoter Prediction

MCPromoter, http://www5.informatik.uni-erlangen.de/HTML/English/Research/Promoter; CoreInspector, http://www.gsf.de/biodv.

Protein Homology

BLOCKS+, http://blocks.fhcrc.org and http:/blocks.fhcrc.org/blocks-bin/getblock.sh?<block name>; GeneWise, http://www.sanger.ac.uk/Software/Wise2/.

Repeat Finders

TRF, http://c3.biomath.mssm.edu/trf.test.html.

ACKNOWLEDGMENTS

We thank all of the participants who submitted their annotations, without which the project would not have been such a success, for their original contributions, their publication, and their patience with the organizers during this very intense project. We also thank the Drosophila Genome Sequencing Center at LBNL, headed by Sue Celniker, for providing such high-quality sequence; the annotation team at the Berkeley Drosophila Genome Sequencing Center and especially Sima Misra, Gerry Rubin, and Michael Ashburner; and the entire *Drosophila* community for producing such a thoroughly studied genomic region. Special thanks go to the independent assessor team, consisting of Michael Ashburner, Peer Bork, Richard Durbin, Roderic Guigó, and Tim Hubbard, who critiqued our evaluation. Thanks goes also to the organizers of ISMB-99 Heidelberg, especially Thomas Lengauer and Reinhard Schneider, for encouraging our tutorial and the tremendous support in the preparation process and during the conference. We also thank Richard Durbin, David Haussler, Tim Hubbard, and Richard Bruskiewich for developing and maintaining the GFF format and their associated tools. Last but not least, a big thank you goes to Gerry Rubin for making the Drosophila Genome Project such a success. This work was supported by NIH grant HG00750.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

Agarwal, P. and D.J. States. 1998. Comparative accuracy of methods for protein sequence similarity search. *Bioinformatics* **14:** 40–47.

Altschul, S.F., W. Gish, W. Miller, E.W. Myers and D.J. Lipman. 1990. Basic local alignment search tool. *J. Mol. Biol.* **215:** 403–410.

Arkhipova, I. R. 1995. Promoter elements in *Drosophila melanogaster* revealed by sequence analysis. *Genetics* **139:** 1359–1369.

Ashburner, M. 2000. A biologist's view of the *Drosophila* genome annotation assessment. *Genome Res.* (this issue).

Ashburner, M., P. Bork, R. Durbin, R. Guigó, and T.J. Hubbard. 1999a. *GASP1 assessment meeting, EMBL*, Heidelberg, Germany.

Ashburner, M., S. Misra, J. Roote, S.E. Lewis, R. Blazej, T. Davis, C. Doyle, R. Galle, R. George, N. Harris et al. 1999b. An exploration of the sequence of a 2.9-Mb region of the genome of drosophila melanogaster. The adh region. *Genetics* **153:** 179–219.

Ashburner, M. et al. 1999c. European *Drosophila* Genome Project (EDGP). http://edgp.ebi.ac.uk/.

Bateman, A., E. Birney, R. Durbin, S.R. Eddy, K.L. Howe, and E.L. Sonnhammer. 2000. The Pfam Protein Families Database. *Nucleic Acids Res.* **28:** 263–266.

Benson, G. 1999. Tandem repeats finder: A program to analyze DNA sequences. *Nucleic Acids Res.* **27:** 573–580.

Besemer, J. and M. Borodovsky. 1999. Heuristic approach to deriving models for gene finding. *Nucleic Acids Res.* **27:** 3911–3920.

Birney, E. 1999. Wise2. http://www.sanger.ac.uk/Software/Wise2/.

Birney, E. and R. Durbin. 1997. Dynamite: A flexible code generating language for dynamic programming methods used in sequence comparison. *Intell. Syst. Mol. Biol.* **5:** 56–64.

———. 2000. Using GeneWise in the *Drosophila* annotation experiment. *Genome Res.* (this issue).

Burge, C. and S. Karlin. 1997. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* **268:** 78–94.

———. 1998. Finding the genes in genomic DNA. *Curr. Opin. Struct. Biol.* **8:** 346–354.

Burset, M. and R. Guigó. 1996. Evaluation of gene structure prediction programs. *Genomics* **34:** 353–367.

Cavin Périer, R., T. Junier, C. Bonnard, and P. Bucher. 1999. The Eukaryotic Promoter Database (EPD): Recent developments. *Nucleic Acids Res.* **27:** 307–309.

Cavin Périer, R., V. Praz, T. Junier, C. Bonnard, and P. Bucher. 2000. The Eukaryotic Promoter Database (EPD). *Nucleic Acids Res.* **28:** 302–303.

Dunbrack, R.L., Jr., D.L. Gerloff, M. Bower, X. Chen, O. Lichtarge, and F.E. Cohen. 1997. Meeting review: The Second meeting on the critical assessment of techniques for protein structure prediction (CASP2), Asilomar, California, December 13–16, 1996. *Folding Design* **2:** R27–R42.

Eeckman, F.H. and R. Durbin. 1995. ACeDB and macace. *Methods Cell Biol.* **48:** 583–605.

Fickett, J.W. and C.S. Tung. 1992. Assessment of protein coding measures. *Nucleic Acids Res.* **20:** 6441–6450.

Fickett, J.W. and A.G. Hatzigeorgiou. 1997. Eukaryotic promoter recognition. *Genome Res.* **7:** 861–878.

Florea, L., G. Hartzell, Z. Zhang, G.M. Rubin, and W. Miller. 1998. A computer program for aligning a cDNA sequence with a genomic DNA sequence. *Genome Res.* **8:** 967–974.

Gaasterland, T. and C.W. Sensen. 1996. MAGPIE: Automated genome interpretation. *Trends Genet.* **12:** 76–78.

Guigó, R., S. Knudsen, N. Drake, and T. Smith. 1992. Prediction of gene structure. *J. Mol. Biol.* **226:** 141–157.

Harris, N.L., G. Helt, S. Misra, and S.E. Lewis. 1999. CloneCurator. http://www.fruitfly.org/displays/CloneCurator.html.

Helt, G., E. Blossom, J. Morris, D. Fineman, S. Cherritz, S. Shaw, and C.L. Harmon. 1999. Neomorphic Genome Software Development Toolkit (NGSDK). Neomorphic, Inc., Berkeley, CA. http://www.neomorphic.com.

Henikoff, S. and J.G. Henikoff. 1994. Protein family classification based on searching a database of blocks. *Genomics* **19:** 97–107.

———. 2000. Genomic sequence annotation based on translated searching of the Blocks+ Database. *Genome Res.* (this issue).

Henikoff, J.G., S. Henikoff, and S. Pietrokovski. 1999a. New features of the Blocks Database servers. *Nucleic Acids Res.* **27:** 226–228.

———. 1999b. Blocks+: A non-redundant database of protein alignment blocks derived from multiple compilations. *Bioinformatics* **15:** 471–479.

Jurka, J. 1998. Repeats in genomic DNA: Mining and meaning. *Curr. Opin. Struct. Biol.* **8:** 333–337.

Krogh, A. 1997. Two methods for improving performance of an HMM and their application for gene finding. *Ismb* **5:** 179–186.

Kulp, D., D. Haussler, M.G. Reese, and F.H. Eeckman. 1997. Integrating database homology in a probabilistic gene structure model. *Pac. Symp. Biocomput.* **2:** 232–244.

Kurtz, S. and C. Schleiermacher. 1999. REPuter: Fast computation of maximal repeats in complete genomes. *Bioinformatics* **15:** 426–427.

Levitt, M. 1997. Competitive assessment of protein fold recognition and alignment accuracy. *Proteins* (Suppl.) **1:** 92–104.

Marcotte, E.M., M. Pellegrini, M.J. Thompson, T.O. Yeates, and D. Eisenberg. 1999. A combined algorithm for genome-wide prediction of protein function. *Nature* **402:** 83–86.

Mott, R. 1997. EST_GENOME: A program to align spliced DNA sequences to unspliced genomic DNA. *Comp. Appl. Biosci.* **13:** 477–478.

Moult, J., T. Hubbard, S.H. Bryant, K. Fidelis, and J.T. Pedersen. 1997. Critical assessment of methods of protein structure prediction (CASP): Round II. *Proteins* (Suppl.) **1:** 2–6.

Moult, J., T. Hubbard, K. Fidelis, and J.T. Pedersen. 1999. Critical assessment of methods of protein structure prediction (CASP): Round III. *Proteins* (Suppl.) **3:** 2–6.

Ohler, U., S. Harbeck, H. Niemann, E. Noth, and M.G. Reese. 1999. Interpolated markov chains for eukaryotic promoter recognition. *Bioinformatics* **15:** 362–369.

Ohler, U., G. Stommer, and S. Harbeck. 2000. Stochastic segment models of eukaroyotic promoter regions. *Pac. Symp. Biocomput.* **5:** 377–388.

Parra, G., E. Blanco, and R. Guigó. 2000. GeneID in *Drosophila. Genome Res.* (this issue).

Pearson, W.R. 1995. Comparison of methods for searching protein sequence databases. *Protein Sci.* **4:** 1145–1160.

Pearson, W.R. and D.J. Lipman. 1988. Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci.* **85:** 2444–2448.

Reese, M.G. 2000. "Genome annotation in *Drosophila melanogaster.*" Ph.D. thesis, University of Hohenheim, Germany.

Reese, M.G., F.H. Eeckman, D. Kulp, and D. Haussler. 1997. Improved splice site detection in Genie. *J. Comput. Biol.* **4:** 311–323.

Reese, M.G., N.L. Harris, G. Hartzell, and S.E. Lewis. 1999. *The 7th conference on Intelligent Systems in Molecular Biology (ISMB'99)*, Heidelberg, Germany, http://www.fruitfly.org/GASP.

Reese, M.G., D. Kulp, H. Tammana, and D. Haussler. 2000. Genie–Gene finding in *Drosophila melanogaster. Genome Res.* (this issue).

Rubin, G.M. 2000. Full-length cDNA project. http://www.fruitfly.org/EST

Rubin, G.M. et al. 1999. Berkeley Drosophia Genome Project (BDGP). http://www.fruitfly.org.

Salamov, A.A. and V.V. Solovyev. 2000. Ab initio gene finding in *Drosophila* genomic DNA. *Genome Res.* (this issue).

Sippl, M.J., P. Lackner, F.S. Domingues, and W.A. Koppensteiner. 1999. An attempt to analyse progress in fold recognition from CASP1 to CASP3. *Proteins* (Suppl.) **3:** 226–230.

Solovyev, V.V., A.A. Salamov, and C.B. Lawrence. 1995. Identification of human gene structure using linear discriminant functions and dynamic programming. *Ismb* **3:** 367–375.

Sonnhammer, E.L., S.R. Eddy, and R. Durbin. 1997. Pfam: A comprehensive database of protein domain families based on seed alignments. *Proteins* **28:** 405–420.

Sonnhammer, E.L., S.R. Eddy, E. Birney, A. Bateman, and R. Durbin. 1998. Pfam: Multiple sequence alignments and HMM-profiles of protein domains. *Nucleic Acids Res.* **26:** 320–322.

Stein, L.D. and J. Thierry-Mieg. 1998. Scriptable access to the *Caenorhabditis elegans* genome sequence and other ACEDB databases. *Genome Res.* **8:** 1308–1315.

Stormo, G.D. 2000. Gene-finding approaches for eukaryotes. *Genome Res. (this issue).*

Uberbacher, E.C. and R.J. Mural. 1991. Locating protein-coding regions in human DNA sequences by a multiple sensor-neural network approach. *Proc. Natl. Acad. Sci.* **88:** 11261–11265.

Zemla, A., C. Venclovas, J. Moult, and K. Fidelis. 1999. Processing and analysis of CASP3 protein structure predictions. *Proteins* (Suppl.) **3:** 22–29.

Received February 9, 2000; accepted in revised form February 29, 2000.

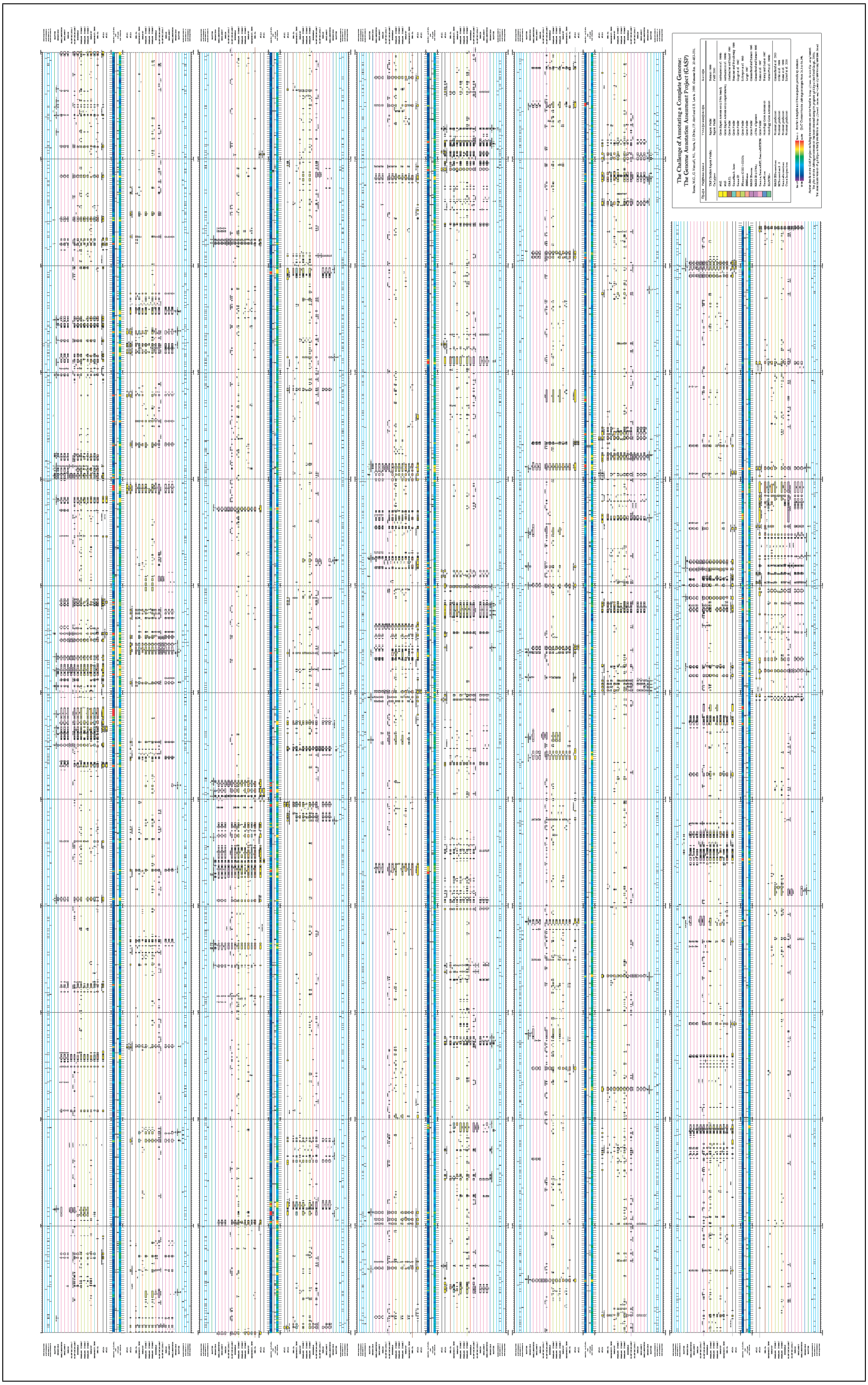

The Challenge of Annotating a Complete Genome:
The Genome Annotation Assessment Project (GASP)

Figure 3.7: ***Drosophila* Genome Annotation Assessment Project.**

3.3.5 Guigó *et al*, *Proc Nat Acad Sci*,100(3):1140–1145, 2003

PubMed Accession:

http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&list_uids=12552088&dopt=Abstract

Journal Abstract:

http://www.pnas.org/cgi/content/abstract/100/3/1140

Supplementary Materials:

http://genome.imim.es/datasets/mouse2002/

Comparison of mouse and human genomes followed by experimental verification yields an estimated 1,019 additional genes

Roderic Guigó*†, Emmanouil T. Dermitzakis†‡, Pankaj Agarwal§, Chris P. Ponting¶, Genís Parra*, Alexandre Reymond‡, Josep F. Abril*, Evan Keibler‖, Robert Lyle‡, Catherine Ucla‡, Stylianos E. Antonarakis‡, and Michael R. Brent‖**

*Research Group in Biomedical Informatics, Institut Municipal d'Investigació Mèdica/Universitat Pompeu Fabra/Centre de Regulació Genòmica, E08003 Barcelona, Catalonia, Spain; ‡Division of Medical Genetics, University of Geneva Medical School and University Hospitals, 1211 Geneva, Switzerland; §GlaxoSmithKline, UW2230, 709 Swedeland Road, King of Prussia, PA 19406; ¶Medical Research Council Functional Genetics Unit, Department of Human Anatomy and Genetics, University of Oxford, South Parks Road, Oxford OX1 3QX, United Kingdom; and ‖Department of Computer Science, Washington University, One Brookings Drive, St. Louis, MO 63130

Communicated by Robert H. Waterston, Washington University School of Medicine, St. Louis, MO, December 11, 2002 (received for review October 21, 2002)

A primary motivation for sequencing the mouse genome was to accelerate the discovery of mammalian genes by using sequence conservation between mouse and human to identify coding exons. Achieving this goal proved challenging because of the large proportion of the mouse and human genomes that is apparently conserved but apparently does not code for protein. We developed a two-stage procedure that exploits the mouse and human genome sequences to produce a set of genes with a much higher rate of experimental verification than previously reported prediction methods. RT-PCR amplification and direct sequencing applied to an initial sample of mouse predictions that do not overlap previously known genes verified the regions flanking one intron in 139 predictions, with verification rates reaching 76%. On average, the confirmed predictions show more restricted expression patterns than the mouse orthologs of known human genes, and two-thirds lack homologs in fish genomes, demonstrating the sensitivity of this dual-genome approach to hard-to-find genes. We verified 112 previously unknown homologs of known proteins, including two homeobox proteins relevant to developmental biology, an aquaporin, and a homolog of dystrophin. We estimate that transcription and splicing can be verified for >1,000 gene predictions identified by this method that do not overlap known genes. This is likely to constitute a significant fraction of the previously unknown, multiexon mammalian genes.

Complete and precise delineation of protein coding genes in mammalian genomes remains a challenging task. To produce a preliminary gene catalog for the draft sequence of the mouse (1), the Mouse Genome Sequencing Consortium relied primarily on the ENSEMBL gene build pipeline (2). ENSEMBL works by (*i*) aligning known mouse cDNAs from REFSEQ (3), RIKEN (4, 5), and SWISSPROT (6, 7) to the genome, (*ii*) aligning known proteins from related mammalian genes to the genome, and (*iii*) using portions of GENSCAN (8) predictions that are supported by experimental evidence (such as ESTs). This conservative approach yielded ≈23,600 genes. However, ENSEMBL cannot predict genes for which there is no preexisting evidence of transcription (1). Furthermore, reliance on known transcripts may lead to a bias against predicting genes that are expressed in a restricted manner or at very low levels.

Before the production of a draft genome sequence for a second mammal, the best available methods for predicting novel mammalian genes were single-genome *de novo* gene-prediction programs, of which GENSCAN (8) is one of the most accurate and most widely used. These programs work by recognizing statistical patterns characteristic of coding sequences, splice signals, and other features in the genome to be annotated. However, they tend to predict many apparently false exons caused by the occurrence of such patterns by chance. With the availability of draft sequences for both the mouse and human genomes, it is now possible to incorporate genomic sequence conservation into *de novo* gene prediction algorithms. However, DNA alignment programs alone are not an effective means of gene prediction because a large fraction of the mouse and human genomes is conserved but does not code for protein.

We developed a procedure that greatly reduces the false-positive rate of *de novo* mammalian gene prediction by exploiting mouse–human conservation in both an initial gene-prediction stage and an enrichment stage. The first stage is to run gene-prediction programs that use genome alignment in combination with statistical patterns in the DNA sequence itself. A number of such programs have been described (9–12). For these experiments, we used SGP2 (13) and TWINSCAN (refs. 14 and 15 and http://genes.cs.wustl.edu), two such programs that we designed for efficient analysis of whole mammalian genomes. TWINSCAN is an independently developed extension of the GENSCAN probability model, whereas SGP2 is an extension of GENEID (16, 17). The probability scores these programs assign to each potential exon are modified by the presence and quality of genome alignments. TWINSCAN uses nucleotide alignment [BLASTN (18), blast.wustl.edu] and has specific models for how alignments modify the scores of coding regions, UTRs, splice sites, and translation initiation and termination signals. SGP2, in contrast, uses translated alignments [TBLASTX (18), blast.wustl.edu] to modify the scores of potential coding regions only. These programs predict many fewer exons than GENSCAN with no reduction in sensitivity to the exons of known genes (13, 14).

The second stage of our procedure is based on the observation that almost all mouse genes have a human counterpart with highly conserved exonic structure (1). We therefore compare all multiexon genes predicted in mouse in the first stage to those predicted in human. Predictions are retained only if the protein predicted in mouse aligns to a human protein predicted by the same program, with at least one predicted intron at the same location (aligned intron, Fig. 1). Predicted single-exon genes are always discarded by this procedure. Although there are many real single-exon genes, it is not currently possible to predict them reliably nor to verify them reliably in a cost-effective, high-throughput procedure.

In this article, we show that our two-stage process yields >1,400 predictions outside the standard annotation of the mouse genome. RT-PCR and direct sequencing of a single exon pair in a sample of these predictions indicates that the majority correspond to real spliced transcripts. Our results also show that this procedure is sensitive to genes that are hard to find by other methods. The combination of these computational and experimental techniques forms a powerful, cost-effective system for expanding experimentally supported genome annotation. This approach is therefore expected to bring the annotation of the mouse and human genomes nearer to closure.

Experimental Procedures

Genome Sequences. The MGSCv3 assembly of the mouse genome described in ref. 1 and the December, 2001 Golden Path assembly

†R.G. and E.T.D. contributed equally to this work.

**To whom correspondence should be addressed. E-mail: brent@cse.wustl.edu.

www.pnas.org/cgi/doi/10.1073/pnas.0337561100

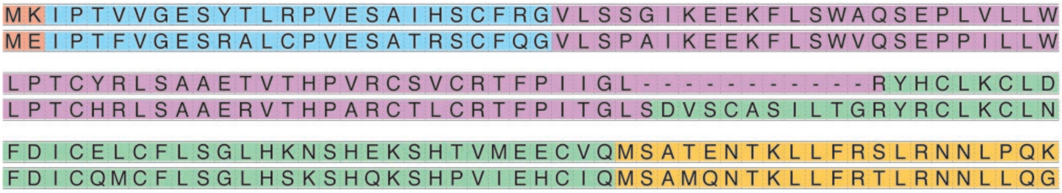

MKIPTVVGESYTLRPVESAIHSCFRGVLSSGIKEEKFLSWAQSEPLVLLW
MEIPTFVGESRALCPVESATRSCFQGVLSPAIKEEKFLSWVQSEPPILLW
LPTCYRLSAAETVTHPVRCSVCRTFPIIGL-----------RYHCLKCLD
LPTCHRLSAAERVTHPARCTLCRTFPITGLSDVSCASILTGRYRCLKCLN
FDICELCFLSGLHKNSHEKSHTVMEECVQMSATENTKLLFRSLRNNLPQK
FDICQMCFLSGLHSKSHQKSHPVIEHCIQMSAMQNTKLLFRTLRNNLLQG

Fig. 1. An example of predictions with aligned introns. RT-PCR positive predicted protein 3B1 (a novel homolog of *Dystrophin*) is aligned with its predicted human ortholog (N-terminal regions shown; *Upper* of each row: mouse, *Lower* of each row: human). Each color indicates one coding exon. Three of four predicted splice boundaries (color boundaries) align perfectly. Any one of these three is sufficient for surviving the enrichment step. Gaps in the alignment (shown as dashes) may indicate mispredicted regions.

of the human genome (National Center for Biotechnology Information Build 28) were downloaded from the University of California (Santa Cruz) genome browser (http://genome.ucsc.edu).

Genome Alignments. TWINSCAN was run on the mouse genome by using BLASTN alignments to the human genome (WU-BLAST, http://blast.wustl.edu). Lowercase masking in the human sequence was first converted to N masking. The result was further masked with NSEG by using default parameters, all Ns were removed, and the sequence was cut into 150-kb database segments. The mouse genome sequence was divided into 1-mb query segments. BLASTN parameters were: M=1 N=−1Q=5 R=1 Z=3000000000 Y=3000000000 B=10000 V=100 W=8 X=20 S=15 S2=15 gapS2=30 lcmask wordmask=seg wordmask=dust topcomboN=3. TWINSCAN was run on the human genome by using separate BLASTN alignments to the mouse genome, which was prepared in the same way except that Ns were not removed before creating the BLAST database.

SGP2 was run on the mouse and human genomes by using a single set of alignments. The masked human genome was cut into 100-kb query segments that were compared with a database of all 100-kb segments of the mouse genome with TBLASTX (WU-BLAST, parameters: B=9000 V=9000 hspmax=500 topcomboN=100 W=5 E=0.01 E2=0.01 Z=3000000000 nogap filter=xnu+seg S2=80). The substitution matrix was BLOSUM62 modified to penalize alignments with stop codons heavily (−500).

Initial Gene Predictions. TWINSCAN was run on 1-mb segments of the mouse and human genomes with target genome parameters identical to the GENSCAN parameters and the 68-set-ortholog conservation parameters (available on request). Note that the TWINSCAN results described in ref. 14 are based on a subsequently developed set of target genome parameters that yields better results than those described here. SGP2 was run on unsegmented mouse and human chromosomes. The REFSEQ genes (which were not tested in the experiments reported here) were incorporated directly into the SGP2 predictions, which improved the predictions outside the REFSEQS slightly by preventing some gene fusion errors. Note that the REFSEQS were not used in generating the SGP2 results described in ref. 13.

Novelty Criteria. Mouse predictions were considered known if they overlapped ENSEMBL predictions or had 95% nucleotide identity to a REFSEQ mRNA or an ENSEMBL-predicted mRNA over at least 100 bp. We used the most inclusive set of ENSEMBL predictions available, based on the complete RIKEN cDNA set without further filtering (1).

Enrichment Procedure. The enrichment procedure was applied separately to predictions of TWINSCAN and SGP2. The protein sequences predicted by each program in human and mouse were compared by using BLASTP (19). For each predicted mouse protein, all predicted human proteins with expect values <1 × 10^{-6} were called homologs. A global protein alignment was produced for the best scoring homologs (up to five) by using T-COFFEE (ref. 39; http://igs-server.cnrs-mrs.fr/∼cnotred/Projects_home_page/t_coffee_home_page.html) with default parameters. Exonic structure was added to the alignments by using EXSTRAL.PL (www1.imim.es/∼rcastelo/exstral.html). When both members of an aligned pair contained an intron at the same coordinate with at least 50% identity over 15 aa on both sides the corresponding mouse prediction was assigned to the "enriched" pool. Predictions with homologs but no aligned intron were assigned to the "similar" pool.

RT-PCR. To test predictions, primers were designed in adjacent exons as described in *Results* and used in RT-PCR of total RNA from 12 normal mouse adult tissues. All procedures were as described (20), except that JumpStart REDTaq ReadyMix (Sigma) and primers from Sigma-Genosys were used.

Additional Details. See supplementary information at www1.imim.es/datasets/mouse2002 for additional details of these procedures.

Results

We applied the two-stage procedure described above to the entire draft mouse and human genome sequences (see *Experimental Procedures*). TWINSCAN predicted 17,271 genes with at least one aligned intron, whereas SGP2 predicted a largely overlapping set of 18,056 genes with at least one aligned intron. These predicted gene sets contain 145,734 exons and 168,492 exons, respectively. Together the two sets overlapped 90% of multiexon ENSEMBL gene predictions.

To estimate a lower bound on the proportion of novel predictions that are transcribed and spliced, we performed a series of RT-PCR amplifications from 12 adult mouse tissues (20). We did not test genes that overlap ENSEMBL predictions nor those that are 95% identical to ENSEMBL predictions or REFSEQ mRNAs over >100 bp or more. Because ENSEMBL was the standard for annotation of the draft mouse genome, we refer to the non-ENSEMBL genes as "novel." A random sample of novel genes predicted by each program and containing at least one aligned intron was tested. Primer pairs were designed in adjacent exons separated by an aligned intron of at least 1,000 bp (Fig. 2). The exon pair to be tested was chosen on the basis of intron length (minimum 1,000 bp), primer design requirements, and *de novo* gene prediction score, with no reference to protein, EST, or cDNA databases. Amplification followed by direct sequencing of the PCR product (Fig. 3) verified the exon pair in 133 unique predicted genes of 214 tested (62%, enriched pool, see Table 1 and www1.imim.es/datasets/mouse2002). Mouse genes predicted by both programs were verified at a much higher rate than those predicted by just one program (76% vs. 27%). Extrapolating from the success rates in Table 1, testing the entire pool of 1,428 enriched predictions in this way is

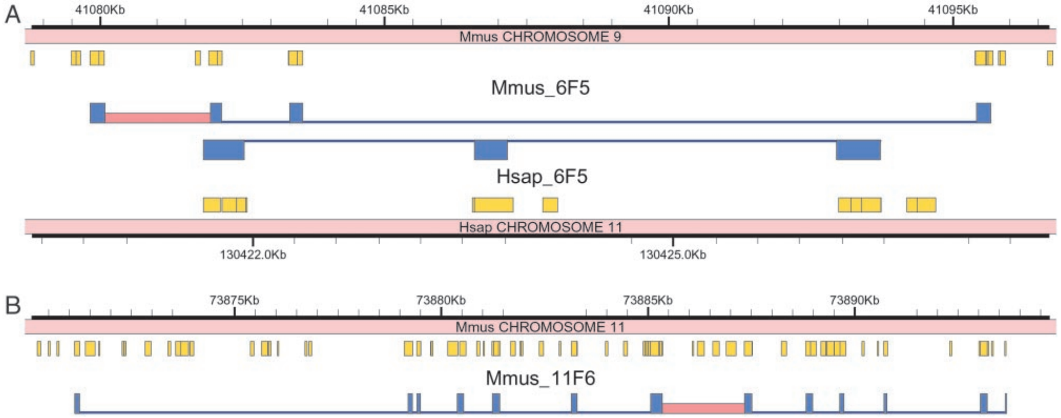


Fig. 2. Two examples of predicted gene structures (blue) with introns verified by RT-PCR from primers located in exons flanking the introns indicated in red. Mouse–human genomic alignments (orange) correlate with predicted exons but do not match them exactly. (*A*) Verified mouse prediction 6F5, a novel homolog of *Drosophila* brain-specific homeobox protein (bsh), with matching human prediction. (*B*) Verified mouse prediction 11F6, a homolog of rat vanilloid receptor type 1-like protein 1. No matching human gene was predicted. A cDNA (GenBank accession no. AF510316) that matches the predicted protein over four protein-coding exons was deposited in GenBank subsequent to our analysis.

expected to yield a total of 788 (±48) predictions with confirmed splices, none of which overlap ENSEMBL predictions.

Considered in isolation, genes predicted by TWINSCAN had a higher verification rate than those predicted by SGP2 (83% vs. 44%), but that difference is skewed by the fact that TWINSCAN predicted fewer exons per gene, and hence its predictions were less likely to overlap ENSEMBL predictions. We corrected for this by clustering overlapping TWINSCAN and SGP2 predictions to ensure that both were counted as positive if either was verified experimentally. For each program, the predictions belonging to a given cluster were counted only once, even if more than one was RT-PCR positive. After this correction, the confirmation rates were much closer (76% for TWINSCAN vs. 62% for SGP2). The results shown in Table 1 include the correction. The TWINSCAN verification rate is similar to the verification rate for genes predicted by both programs because the exons predicted by TWINSCAN are largely a subset of those predicted by SGP2.

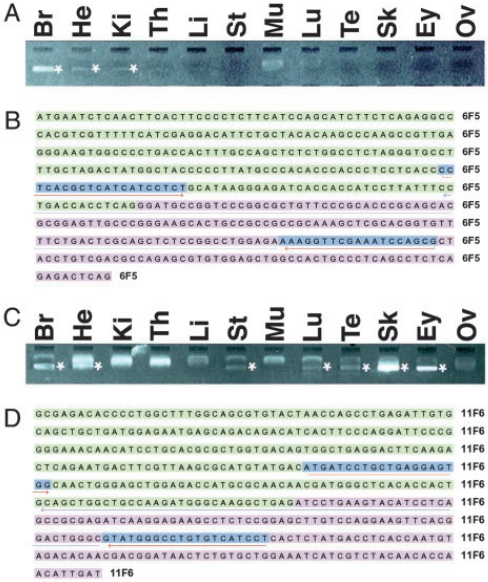

A
Br
He
Ki
Th
Li
St
Mu
Lu
Te
Sk
Ey
Ov
*
B
ATGAATCTCAACTTCACTTCCCCTCTTCATCCAGCATCTTCTCAGAGGCC 6F5
CACGTCGTTTTTCATCGAGGACATTCTGCTACACAAGCCCAAGCCGTTGA 6F5
GGGAAGTGGCCCCTGACCACTTTGCCAGCTCTCTGGCCTCTAGGGTGCCT 6F5
TTGCTAGACTATGGCTACCCCCTTATGCCCACACCCACCCTCCTCACCCC 6F5
TCACGCTCATCATCCTCTGCATAAGGGAGATCACCACCATCCTTATTTCC 6F5
TGACCACCTCAGGGATGCCGGTCCCGGCGCTGTTCCCGCACCCGCAGCAC 6F5
GCGGAGTTGCCCGGGAAGCACTGCCGCCGCCGCAAAGCTCGCACGGTGTT 6F5
TTCTGACTCGCAGCTCTCCGGCCTGGAGAAAAGGTTCGAAATCCAGCGCT 6F5
ACCTGTCGACGCCAGAGCGTGTGGAGCTGGCCACTGCCCTCAGCCTCTCA 6F5
GAGACTCAG 6F5
C
Br
He
Ki
Th
Li
St
Mu
Lu
Te
Sk
Ey
Ov
D
GCGAGACACCCCTGGCTTTGGCAGCGTGTACTAACCAGCCTGAGATTGTG 11F6
CAGCTGCTGATGGAGAATGAGCAGACAGACATCACTTCCCAGGATTCCCG 11F6
GGGAAACAACATCCTGCACGCGCTGGTGACAGTGGCTGAGGACTTCAAGA 11F6
CTCAGAATGACTTCGTTAAGCGCATGTATGACATGATCCTGCTGAGGAGT 11F6
GGCAACTGGGAGCTGGAGACCATGCGCAACAACGATGGGCTCACACCACT 11F6
GCAGCTGGCTGCCAAGATGGGCAAGGCTGAGATCCTGAAGTACATCCTCA 11F6
GCCGCGAGATCAAGGAGAAGCCTCTCCGGAGCTTGTCCAGGAAGTTCACG 11F6
GACTGGGCGTATGGGCCTGTGTCATCCTCACTCTATGACCTCACCAATGT 11F6
AGACACAACGACGGATAACTCTGTGCTGGAAATCATCGTCTACAACACCA 11F6
ACATTGAT 11F6

Fig. 3. Verification of gene predictions by RT-PCR analysis. (*A* and *B*) Test of prediction 6F5, a homolog of *Drosophila* brain-specific homeobox protein (bsh). (*C* and *D*) Test of prediction 11F6, a homolog of rat vanilloid receptor type 1-like protein. Gel analysis of amplimers (*) with the source of the cDNA pool indicated above is shown in *A* and *C*. Primers (blue) and the region to which the amplimer sequence aligned (underlining) are shown in *B* and *D*. The indicated forward primers were used to generate the amplimer sequences (brain amplimer, *B*; skin amplimer, *D*). Br, brain; Ey, eye; He, heart; Ki, kidney; Li, liver; Lu, lung; Mu, muscle; Ov, ovary; Sk, skin; St, stomach; Te, testis; Th, thymus.

Before the enrichment procedure, the combined predictions of SGP2 and TWINSCAN overlap 98% of multiexon ENSEMBL genes, as compared with 90% for the enriched pool. This finding suggests that the enrichment procedure reduces sensitivity by a small but noticeable degree. To investigate the potential loss of sensitivity further, we applied the same RT-PCR procedure to two samples of gene predictions that were excluded by the enrichment criterion and did not overlap ENSEMBL predictions. One sample had one or more regions of strong similarity to a predicted human gene but did not satisfy the aligned intron criterion (similar pool) whereas the other lacked any strong similarity to a human prediction by the same program (other pool). The verification rates for the similar and other pools were 25% and 20%, respectively, for genes predicted by both programs, and 0% and 2%, respectively, for genes predicted by only one program (Table 1 and www1.imim.es/datasets/mouse2002). This finding shows that the enrichment procedure increases specificity greatly and, consistent with the ENSEMBL overlap analysis, reduces sensitivity only slightly. If all predictions in the similar and other pools were tested the expected numbers of successes are 126 (±105) and 105 (±83), respectively, with the large standard errors resulting from the small number of successful amplifications in these pools.

As a control, we also tested 113 predictions from the enriched pool that did overlap ENSEMBL predictions. In 66 of the predictions the splice boundary we tested was predicted identically in ENSEMBL, and 64 of these tests (97%) were positive. In 47 of the predictions the splice boundary we tested was not predicted identically in ENSEMBL, and 21 of these tests (45%) were positive,

Table 1. Predicted novel gene sets and RT-PCR verification rates

Pool	Programs*	No. of predictions	No. tested	No. positive	Success rate, %	Expected successes	Standard error
Enriched†	Both	827	154	117	75.97	628	
	One	601	60	16	26.67	160	
	Total	1,428	214	133	62.15	788	48
Similar‡	Both	505	16	4	25.00	126	
	One	1,620	22	0	0.00	0	
	Total	2,125	38	4	10.53	126	105
Other§	Both	234	5	1	20.00	46	
	One	3,425	58	1	1.72	59	
	Total	3,659	63	2	3.17	105	83
All	Total	7,212	315	139	N/A	1,019	

N/A, not applicable.

*Both, Genes predicted at least partially by both TWINSCAN and SGP2 programs. One, Genes predicted by one program that are not overlapped by predictions of the other program. N/A, not applicable.

†Mouse gene predictions containing an intron whose flanking exonic regions align with flanking exonic regions predicted by the same program in human.

‡Mouse gene predictions that fail the enrichment step but show regions of strong similarity to a gene predicted by the same program in human.

§Mouse gene predictions without regions of strong similarity to any gene predicted by the same program in human.

despite the fact that ENSEMBL predictions are based on transcript evidence. This verification rate may reflect alternative splices identified by our method but not by ENSEMBL.

To determine whether tissue-restricted expression could explain the absence of the predictions we verified from the transcript-based annotation, we compared the expression patterns of our RT-PCR positive predictions to those of the complete set of mouse orthologs of genes mapping to human chromosome 21 (Hsa21). These genes were chosen for comparison because they had been previously subjected to the same protocol with the same cDNA pools in the same laboratory (20). Our verified novel gene predictions showed a significantly more restricted pattern of expression (Fig. 4*A*). The mean number of tissues for our positive predictions was 6.3, and 33% of the positive predictions showed expression in three or fewer tissues; the corresponding numbers for the mouse orthologs of human chromosome 21 genes are 8.2 tissues on average and 14% showing expression in three or fewer tissues. This difference in expression specificity was statistically significant (ANOVA, $F = 23.22$, df = 1, $P < 0.001$).

To determine whether prediction of pseudogenes by our method could explain some of the RT-PCR negatives, we computed the ratio of nonsynonymous to synonymous substitution rates (K_A/K_S) (21) for the subset of tested mouse predictions with unique putative human orthologs (Fig. 4*B*). The mean for PCR-positive predictions was 0.29 whereas for PCR-negative predictions it was 0.72. The difference was statistically significant (ANOVA, $F = 34.86$, df = 1, $P < 0.001$), suggesting that (*i*) some of the negative predictions may be pseudogenes, and (*ii*) K_A/K_S can be efficiently incorporated in the enrichment protocol to increase specificity (22).

Among the predictions with confirmed splices, 112 had significant homology to known genes and/or domains. A few of these genes, which were not represented in databases at the beginning of our gene survey, were submitted to databases and/or published in the literature in the intervening months. For example, we correctly predicted the first four protein coding exons of *TRPV3*, a heat-sensitive TRP channel in keratinocytes (23), and both exons of *RLN3* (*preprorelaxin 3*), an insulin-like prohormone (24). The verified predictions with the most notable homologies are shown in Table 2, including a novel homolog of dystrophin that is discussed in the mouse genome paper (1). Table 2 includes two noncanonical homeobox genes, one that is most similar to fruitfly brain-specific homeobox protein (Figs. 2 and 3 *A* and *B*) (25) and another that is a Not-class homeobox, likely to be involved in notochord development (26). Four predicted genes were found to be expressed in the brain and are likely to have neuronal functions, including one paralog each of: *Nna1*, which is expressed in regenerating motor neurons (27); an *N*-acetylated-α-linked-acidic dipeptidase, which hydrolyses the neuropeptide *N*-acetyl-aspartyl-glutamate to terminate its neurotransmitter activity (28); a novel γ-aminobutyric acid type B receptor, which regulates neurotransmitter release (29); and an Ent2-like nucleoside transporter, which modulates neurotransmission by altering adenosine concentrations (30). Other verified genes are likely to be important in muscle contraction (myosin light chain kinase homolog), degradation of cell cycle proteins (fizzy/CDC20 homolog), Wnt-dependent vertebrate development (Dapper/frodo homolog), and solute and steroid transport in the liver (solute transporter β). Homologs of two further genes predicted in our studies are associated with disease. *ATP10C*, an aminophospholipid translocase, is absent from Angelman syndrome patients with imprinting mutations (31), and *otoferlin,* which is mutated in a nonsyndromic form of deafness (32).

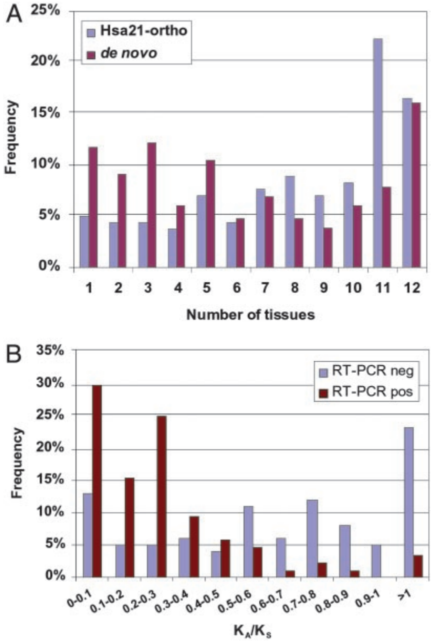


Fig. 4. Characteristics of verified predictions. (*A*) Expression specificity. Percentages of RT-PCR positive *de novo* predictions (red) and Hsa21 mouse orthologs (blue) expressed in 1–12 tissues, tested in the same cDNA pools. (*B*) Distributions of the ratio of nonsynonymous to synonymous substitution rate (K_A/K_S) in 83 RT-PCR positive (red) vs. 98 RT-PCR negative (blue) mouse predictions with reciprocal best BLAST matches among the human predictions.

Table 2. Novel mouse genes, their tissue expression, and their homologs

Code	B	H	K	Y	V	S	M	L	T	K	E	O	%Id	Ln	Homology
3B1								+	+				38	134	Dystrophin-like; with ZZ domain
3B3				+		+		+	+	+			25	184	Novel aquaporin; similar to *Drosophila* CG12251
3C3			+		+			+	+		+		25	260	TEP1 (telomerase associated); probable ATPase
3C5									+		+		47	198	Voltage-dependent calcium channel γ subunit
4B3			+			+			+				34	74	IFN-induced/fragilis transmembrane family
4C6		+				+		+	+	+			30	134	IL-22-binding protein CRF2-10
4G4	+								+	+	+		64	109	Nna1p, nuclear ATP/GTP-binding protein
5B5						+			+	+			43	111	Likely aminophospholipid flippase (transporting ATPase)
1E3	+			+	+			+			+		40	106	*N*-acetylated-α-linked-acidic dipeptidase (NAALADase)
6C4								+	+				42	117	Not-type homeobox; poss. involved in notochord development
6F5	+	+	+										66	102	*Drosophila* brain-specific homeobox protein (bsh)
11F2	+					+		+	+	+			29	216	Human γ-aminobutyric acid type B receptor 2, neurotransmitter release regulator
5A2			+		+	+			+				41	36	Skate liver organic solute transporter β
11B6				+				+			+		55	116	IFN-activatable protein 203; nuclear protein
12B3	+			+	+	+		+	+	+	+		25	229	Fatty acid desaturase; maintains membrane integrity
11F6	+	+				+		+	+	+	+		44	494	Rat vanilloid receptor type 1 like protein 1
12E3								+	+				52	175	Fizzy/CDC20; modulates degradation of cell-cycle proteins
12F1		+				+	+	+	+				43	355	Otoferlin (mutated in DFNB9, nonsyndromic deafness)
12H1	+	+							+				45	116	Fruitfly additional sex combs; a Polycomb group protein
12C4	+							+			+		43	133	*Caenorhabditis elegans* C15C8.2; single-minded-like; HLH and PAS domains
12D2						+							41	397	Cytosolic phospholipase A2, group IVB
12A5	+												38	415	Fruitfly GH15686p; Ent2-like nucleoside transporter
12E5	+			+				+				+	32	111	Relaxin 3 preproprotein; prohormone of the insulin family
11A1			+	+	+		+					+	89	75	Mouse BET3, involved in ER to Golgi transport
11A2	+	+						+	+		+	+	70	207	Vacuolar ATP synthase subunit S1
11B2							+	+	+	+	+	+	54	271	Myosin light chain kinase, skeletal muscle
11G2	+		+	+	+	+		+	+	+	+	+	36	179	Dapper/frodo (transduces Wnt signals by interacting with Dsh)

Code, Coding name of tested gene model. B, brain; H, heart; K, kidney; Y, thymus; V, liver; S, stomach; M, muscle; L, lung; T, testis; K, skin; E, eye; O, ovary. %Id, Percentage amino acid identity. Ln, Number of amino acids in the local alignment between the prediction and the homolog.

Discussion

We have demonstrated a remarkably efficient mammalian gene discovery system. This system exploits the draft mouse and human genome sequences in both an initial gene-prediction stage and an enrichment stage. The first stage consists of SGP2 and TWINSCAN, gene-prediction programs that use genome alignment in combination with statistical patterns in the DNA sequence. We have shown elsewhere that both programs have greater sensitivity and specificity than single-genome *de novo* predictors, such as GENSCAN (13, 14). In this article, we have demonstrated the effectiveness of the enrichment stage, in which predictions are retained only if the protein predicted in mouse aligns to a human protein predicted by the same program, with at least one predicted intron at the same location (aligned intron, Fig. 1). In our pool of predictions, the aligned intron filter is expected to eliminate 24 times more RT-PCR negatives than RT-PCR positives. This enrichment procedure can be applied to predictions from any program.

Our goal was to develop a low-cost, high-throughput system for finding and verifying coding regions that are missed by annotation systems that require existing transcript evidence. ENSEMBL was chosen as the representative of such systems because the Mouse Genome Sequencing Consortium judged it to be the most suitable tool for timely, cost-effective, reliable annotation of the mouse genome sequence. Thus, we evaluated our system by investigating genes that do not overlap ENSEMBL predictions. Our system is not designed to find genes that would be missed by expert manual annotators, who can effectively integrate information such as the predictions of GENSCAN (8) and GENOMESCAN (33), percent-identity plots (34), comparison to fish genomes (35, 36), alignment of weakly homologous proteins, and alignment of EST sequences. As a result, we did not exclude gene predictions from our evaluation based on these indicators.

Our two-stage system identified a highly reliable pool of 827 predicted genes not overlapping the standard annotation, of which we tested 154 for expression by using RT-PCR and direct sequencing. Primers designed for a single pair of adjacent exons in each predicted gene yielded a spliced PCR product whose sequence closely matched that of the predicted exons in 76% of these tests. In the only other published report of high-throughput verification of gene predictions of which we are aware, 14% of predictions not overlapping the standard annotation yielded spliced products (37). These numbers cannot be compared directly because of differences in the sampling criteria, but the magnitude of the difference suggests our method provides new levels of efficiency in experimental confirmation of genes outside the standard annotation set.

The sensitivity of our method also appears to be high. Predictions in our enriched pool overlap 90% of multiexon genes predicted by ENSEMBL. However, it has been estimated that >4,000 ENSEMBL predictions comprising 12,000 predicted exons are in fact pseudogenes (1). Although the precise number of multiexon pseudogenes in the ENSEMBL annotation is unknown, this estimate suggests that our enriched pool may overlap a much larger fraction of the functional genes identified by ENSEMBL. Further, RT-PCR tests of TWINSCAN and SGP2 predictions outside the enriched pool indicate that a relatively small number of these predictions are transcribed and spliced in the 12 tissues tested. Thus, the enrichment procedure is sensitive to both ENSEMBL predictions and verifiable predictions by TWINSCAN and SGP2.

Using our system, we confirmed one intron of 139 predicted genes that do not overlap any gene in the standard mouse genome annotation (1). Ninety-two of the RT-PCR positive introns (66%) did not align to any mouse EST, and these might have posed difficulties even for human annotators. Furthermore, seven of the RT-PCR negative introns (4%) did align to mouse ESTs and six of these were in the enriched pool, suggesting that the true percentage of transcribed and spliced predictions in this pool may be even higher than the RT-PCR positive percentage.

Among RT-PCR positive predictions, 24 had homologies to known proteins that we found particularly interesting (Table 2). The positive identification of these homologs is expected to impact numerous research programs devoted to genes of developmental and medical importance. In general, these genes were probably missed in the ENSEMBL annotation because the length and percent identity of the homologies were not sufficient to support a protein-based gene prediction (Table 2). In many cases, such as the predicted homolog of a brain-specific homeobox protein, the expression patterns we found were consistent with what would be expected from the function of the known homolog (Fig. 3*A* and *B*).

The confirmed 139 genes also showed a relatively restricted expression pattern, on average. Because all mouse orthologs of genes on human chromosome 21 had already been tested by using the same experimental protocol and the same cDNA pools, we were able to directly compare expression patterns. To the extent that the known genes on chromosome 21 are no more tissue specific than the complete set of known genes, the results (Fig. 4) suggest that our system may be particularly sensitive to genes with tissue-restricted expression. Qualitatively similar restricted expression patterns were reported for novel GENSCAN predictions on chromosome 22 (37), lending further support to the value of *de novo* prediction for identifying genes with tissue-restricted expression.

Of the RT-PCR positive novel predictions, only 33% have identifiable homologs in the sequenced fish (*Fugu*/*Tetraodon*/zebrafish) genomes. Comparing this finding to the recent estimate that three-quarters of all human genes can be recognized in the *Fugu* genome (36) suggests that our system may be particularly sensitive to genes that are not ubiquitous in the vertebrate lineage. Genes with relatively restricted expression patterns and species distribution can be difficult to find by using transcript-based methods like GENEWISE (38) and compact-genome methods like EXOFISH (35), but they appear to be tractable for our system.

Extrapolating from the success rates in all categories, the expected total number of gene predictions that could be successfully RT-PCR amplified in the cDNA pools we tested is 1,019 (Table 1), adding ≈5% to the number of functional mouse genes identified by ENSEMBL (1). The number of distinct genes verifiable in this way may be slightly smaller, because the effect of fragmentation in ENSEMBL and in our predictions is not readily testable. However, the number of predictions that are transcribed and spliced is likely to be >1,019, because (*i*) we tested only one exon pair from each prediction and (*ii*) we used only 12 adult mouse tissues (20).

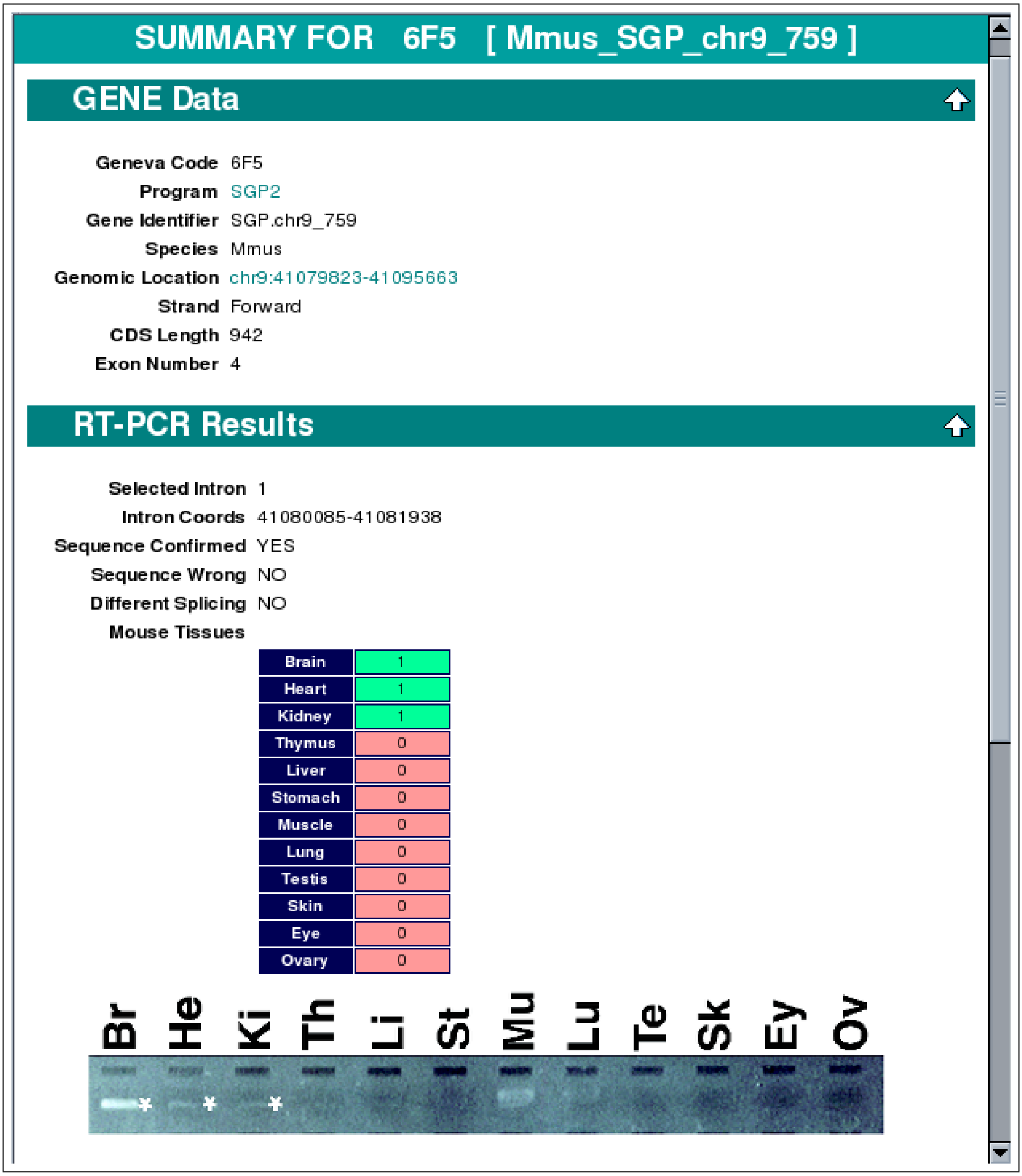
The relatively low success rate in the pools failing the enrichment step suggests that the number of real, multiexon genes whose existence has been predicted but not yet confirmed is in the range of 1,000–2,000 (including those predictions in the enriched pool that have not been confirmed). Because we have used only two prediction programs, TWINSCAN and SGP2, it is possible that other programs might yield a large additional set of predictions that pass the enrichment step. However, GENSCAN yields only 49 additional predictions that pass enrichment and novelty criteria and do not overlap the 1,428 "aligned intron" novel predictions from TWINSCAN and SGP2 (3%). These 49 are worth testing, and adding more prediction programs will yield at least a few more predictions with aligned introns. Nonetheless, the data presented here suggest that the 1,428 predictions in the enriched pool may overlap a significant fraction of the previously unannotated, multiexon mouse genes.

Using the draft sequences of the mouse and human genomes, we have developed a cost-effective, high-throughput system for predicting genes and verifying the existence of corresponding spliced transcripts. Applying this system to the entire mouse genome, we showed that an automated system can produce a large set of experimentally supported mammalian gene predictions outside the standard annotation. Further, the average cost per verified exon pair is less than two primer pairs and sequencing reactions. We expect that testing the remaining predictions in the enriched pool will locate most multiexon mouse genes that are currently unannotated, bringing us significantly closer to identification of the complete mammalian gene set.

As more mammalian genomes are sequenced, the need for experimentally validated high-throughput annotation will continue to grow, as will the data available for methods such as ours. Using the sequences of more genomes, it may be possible to extend this approach to single-exon and lineage-specific genes. In combination with methods like ENSEMBL and refinement by expert annotators, these developments may bring complete, experimentally supported genome annotation within reach.

We are grateful to the Mouse Genome Sequencing Consortium for providing the mouse genome sequence as well as support throughout the analysis process. We are particularly grateful to Eric Lander, Robert Waterston, Ewan Birney, Adam Felsenfeld, and Ross Hardison for advice and encouragement. Thanks are also due to Marc Vidal, Lior Pachter, Kerstin Lindblad-Toh, and Gwen Acton for participation in pilot experiments and Tamara Doering for helpful comments on the manuscript. Research at Institut Municipal d'Investigació Mèdica/Universitat Pompeu Fabra/Centre de Regulació Genòmica is supported by a grant from the Spanish Plan Nacional de Investigacion y Desarrollo. J.F.A. is supported by a fellowship from the Instituto de Salud Carlos III. The Division of Medical Genetics is supported by the Swiss National Science Foundation, National Centres of Competence in Research Frontiers in Genetics, and the Childcare and J. Lejeune Foundations. Research at Washington University was supported by Grant DBI-0091270 from the National Science Foundation (to M.R.B.) and Grant HG02278 from the National Institutes of Health (to M.R.B.).

1. Mouse Genome Sequencing Consortium (2002) *Nature* **420,** 520–562.
2. Hubbard, T., Barker, D., Birney, E., Cameron, G., Chen, Y., Clark, L., Cox, T., Cuff, J., Curwen, V., Down, T., *et al*. (2002) *Nucleic Acids Res.* **30,** 38–41.
3. Pruitt, K. D. & Maglott, D. R. (2001) *Nucleic Acids Res.* **29,** 137–140.
4. Kawai, J., Shinagawa, A., Shibata, K., Yoshino, M., Itoh, M., Ishii, Y., Arakawa, T., Hara, A., Fukunishi, Y., Konno, H., *et al*. (2001) *Nature* **409,** 685–690.
5. The FANTOM Consortium and The RIKEN Genome Exploration Research Group Phase II Team (2002) *Nature* **420,** 563–571.
6. Bairoch, A. & Apweiler, R. (2000) *Nucleic Acids Res.* **28,** 45–48.
7. Gasteiger, E., Jung, E. & Bairoch, A. (2001) *Curr. Issues Mol. Biol.* **3,** 47–55.
8. Burge, C. & Karlin, S. (1997) *J. Mol. Biol.* **268,** 78–94.
9. Wiehe, T., Gebauer-Jung, S., Mitchell-Olds, T. & Guigó, R. (2001) *Genome Res.* **11,** 1574–1583.
10. Pachter, L., Alexandersson, M. & Cawley, S. (2002) *J. Comput. Biol.* **9,** 389–399.
11. Batzoglou, S., Pachter, L., Mesirov, J. P., Berger, B. & Lander, E. S. (2000) *Genome Res.* **10,** 950–958.
12. Bafna, V. & Huson, D. H. (2000) *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **8,** 3–12.
13. Parra, G., Agarwal, P., Abril, J. F., Wiehe, T., Fickett, J. W. & Guigó, R. (2003) *Genome Res.* **13,** 108–117.
14. Flicek, P., Keibler, E., Hu, P., Korf, I. & Brent, M. R. (2003) *Genome Res.* **13,** 46–54.
15. Korf, I., Flicek, P., Duan, D. & Brent, M. R. (2001) *Bioinformatics* **17,** Suppl. 1., S140–S148.
16. Parra, G., Blanco, E. & Guigó, R. (2000) *Genome Res.* **10,** 511–515.
17. Guigó, R., Knudsen, S., Drake, N. & Smith, T. (1992) *J. Mol. Biol.* **226,** 141–157.
18. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990) *J. Mol. Biol.* **215,** 403–410.
19. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997) *Nucleic Acids Res.* **25,** 3389–3402.
20. Reymond, A., Marigo, V., Yaylaoglu, M. B., Leoni, A., Ucla, C., Scamuffa, N., Caccioppoli, C., Dermitzakis, E. T., Lyle, R., Banfi, S., *et al*. (2002) *Nature* **420,** 582–586.
21. Hughes, A. L. & Nei, M. (1988) *Nature* **335,** 167–170.
22. Nekrutenko, A., Makova, K. D. & Li, W. H. (2002) *Genome Res.* **12,** 198–202.
23. Peier, A. M., Reeve, A. J., Andersson, D. A., Moqrich, A., Earley, T. J., Hergarden, A. C., Story, G. M., Colley, S., Hogenesch, J. B., McIntyre, P., *et al*. (2002) *Science* **296,** 2046–2049.
24. Bathgate, R. A., Samuel, C. S., Burazin, T. C., Layfield, S., Claasz, A. A., Reytomas, I. G., Dawson, N. F., Zhao, C., Bond, C., Summers, R. J., *et al*. (2002) *J. Biol. Chem.* **277,** 1148–1157.
25. Jones, B. & McGinnis, W. (1993) *Development (Cambridge, U.K.)* **117,** 793–806.
26. Talbot, W. S., Trevarrow, B., Halpern, M. E., Melby, A. E., Farr, G., Postlethwait, J. H., Jowett, T., Kimmel, C. B. & Kimelman, D. (1995) *Nature* **378,** 150–157.
27. Harris, A., Morgan, J. I., Pecot, M., Soumare, A., Osborne, A. & Soares, H. D. (2000) *Mol. Cell. Neurosci.* **16,** 578–596.
28. Pangalos, M. N., Neefs, J. M., Somers, M., Verhasselt, P., Bekkers, M., van der Helm, L., Fraiponts, E., Ashton, D. & Gordon, R. D. (1999) *J. Biol. Chem.* **274,** 8470–8483.
29. Billinton, A., Ige, A. O., Bolam, J. P., White, J. H., Marshall, F. H. & Emson, P. C. (2001) *Trends Neurosci.* **24,** 277–282.
30. Crawford, C. R., Patel, D. H., Naeve, C. & Belt, J. A. (1998) *J. Biol. Chem.* **273,** 5288–5293.
31. Meguro, M., Kashiwagi, A., Mitsuya, K., Nakao, M., Kondo, I., Saitoh, S. & Oshimura, M. (2001) *Nat. Genet.* **28,** 19–20.
32. Yasunaga, S., Grati, M., Cohen-Salmon, M., El-Amraoui, A., Mustapha, M., Salem, N., El-Zir, E., Loiselet, J. & Petit, C. (1999) *Nat. Genet.* **21,** 363–369.
33. Yeh, R. F., Lim, L. P. & Burge, C. B. (2001) *Genome Res.* **11,** 803–816.
34. Schwartz, S., Zhang, Z., Frazer, K. A., Smit, A., Riemer, C., Bouck, J., Gibbs, R., Hardison, R. & Miller, W. (2000) *Genome Res.* **10,** 577–586.
35. Roest Crollius, H., Jaillon, O., Bernot, A., Dasilva, C., Bouneau, L., Fischer, C., Fizames, C., Wincker, P., Brottier, P., Quetier, F., *et al*. (2000) *Nat. Genet.* **25,** 235–238.
36. Aparicio, S., Chapman, J., Stupka, E., Putnam, N., Chia, J. M., Dehal, P., Christoffels, A., Rash, S., Hoon, S., Smit, A., *et al*. (2002) *Science* **297,** 1301–1310.
37. Das, M., Burge, C. B., Park, E., Colinas, J. & Pelletier, J. (2001) *Genomics* **77,** 71–78.
38. Birney, E. & Durbin, R. (2000) *Genome Res.* **10,** 547–548.
39. Notre dame, C., Higgins, D. G. & Heringa, J. (2000) *J. Mol. Biol.* **302,** 205–217.



SUMMARY FOR 6F5 [Mmus_SGP_chr9_759]

GENE Data

Geneva Code	6F5
Program	SGP2
Gene Identifier	SGP.chr9_759
Species	Mmus
Genomic Location	chr9:41079823-41095663
Strand	Forward
CDS Length	942
Exon Number	4

RT-PCR Results

Selected Intron	1
Intron Coords	41080085-41081938
Sequence Confirmed	YES
Sequence Wrong	NO
Different Splicing	NO
Mouse Tissues	

Brain	1
Heart	1
Kidney	1
Thymus	0
Liver	0
Stomach	0
Muscle	0
Lung	0
Testis	0
Skin	0
Eye	0
Ovary	0

Figure 3.10: **A web server to display RT-PCR results over predicted genes.** A small database containing all the 476 genes that were submitted for the RT-PCR validation test was provided as supplementary materials for Guigó *et al.* [2003, see also page 215 on web glossary]. That pool of genes was filtered out from the gene predictions by `SGP2` and `Twinscan` on the mouse genome by exploiting conservation in mouse-human exonic structure.

Chapter 4

Sequence features of Eukaryotic Genes

The human mind has first to construct forms,
independently, before we can find them in things.
—Albert Einstein

Most genes in higher eukaryotes are interrupted by non-coding sequences (introns) that must be precisely excised from pre-messenger RNA (pre-mRNA) molecules to yield mature, functional mRNAs. In those organisms, splicing introduces an additional level of decoding on the sequence of the primary RNA transcript, prior to translation. The genetic code is essentially deterministic. Within a given species, a given triplet in the mRNA sequence results always in the same amino acid. In contrast, the splicing code is inherently stochastic. The probability of a splicing sequence in the primary transcript to participate in the definition of an intron boundary ranges from zero to one, and is conditioned to very many different factors.

The unexpected discovery in 1977 of split genes in the adenovirus 2 (*Ad2*) mRNAs [Berget *et al.*, 1977; Chow *et al.*, 1977], started an amazing scientific endeavour. In this chapter we start with an overview of the current knowledge about the splicing process at molecular level. Then we report a comparative computational analysis of orthologous splice sites of four vertebrate genomes.

4.1 The Molecular Basis of Splicing

A typical mammalian gene contains nine introns and spans about 30 kb. An average intron is over 3000 bp long, while an average exon is only about 150 bp [Lander *et al.*, 2001]. It has been known for a long time that intron removal and the ligation of flanking sequences (exons) occurs through two sequential trans-esterification reactions that are carried out by a multicomponent complex that is known as the spliceosome (see Figure 4.1).

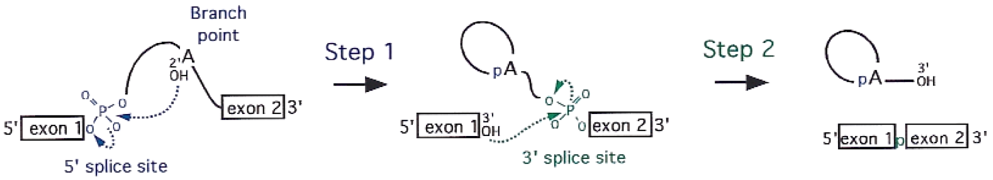


Figure 4.1: **The splicing reaction at the biochemical level.** The pre-mRNA splicing reaction consists of two phosphoryl-transfer steps. In the first step, the 5′ phosphate of the intron (at the 5′ splice site) is attacked by a 2′ hydroxyl specified within the intron (from the adenosine in the branch point). In a second step, the 3′ phosphate of the intron (at the 3′ splice site) is attacked by the 3′ hydroxyl of the cleaved 5′ exon. The final products are ligated exons and the excised intron in a branched form also known as lariat. Adapted from Collins and Guthrie [2000].

Most introns have common consensus sequences near their 5′ and 3′ ends that are recognized by spliceosomal components and are required for spliceosome formation. The assembly of a spliceosome on a pre-mRNA is an ordered process that involves five small nuclear ribonucleoprotein particles (snRNPs: U1, U2, U4, U5 and U6), as well as an array of protein factors. Catalysis of the splicing reaction proceeds by coordinated series of RNA-RNA, RNA-protein and protein-protein interactions, which lead to exon ligation and release of the intron lariat [Patel and Steitz, 2003].

4.1.1 U2 versus U12 splice sites

The first intron sequences ever characterized revealed highly conserved dinucleotides at the 5′ and 3′ termini (GT and AG, respectively). They were later found to be parts of longer consensus sequences at the 5′ and 3′ splice sites (such as those represented in Figure 1.3 on page 4 and those shown in Figure 4.12 on page 130 (Figure 1 on page 112 of Abril *et al.* 2005). The canonical splice site consensus sequence was first catalogued by Mount [1982] an later refined with more data by Senapathy *et al.* [1990]. The presence of non-consensus splice sites was first recognized in Jackson [1991], but it was not proposed that there was a distinct minor class of introns until the works of Hall and Padgett [1994]. They noted that four introns shared unusual consensus sequences, and predicted that their excision was mediated by a distinct spliceosome that involved low-abundance snRNPs (less than 10^4 copies per cell), U11 and U12 [Montzka and Steitz, 1988], for which no function had been described at that time. Indeed, U11 and U12 have base-pairing potential with the 5′ splice-site and branch-site sequences, whereas their secondary structures mimic those of U1 and U2, respectively (Figure 4.2).

Because these new introns had AT and AC termini, which deviates from the nearly invariant GT-AG rule, they were initially named AT-AC introns. However, more extensive genomic database surveys revealed that AT-AC termini are not a defining feature of the minor class introns [Dietrich *et al.*, 1997; Sharp and Burge, 1997; Wu and Krainer, 1997]. In fact, most minor-class introns have canonical GT-AG termini and, very rarely, major-class introns have AT-AC termini [Sharp and Burge, 1997]. An analysis of canonical and non-canonical splice sites in mammalian genomes [Burset *et al.*, 2000] estimated the occurrence of different splice site termini: GT-AG (99.20%), CG-AG (0.62%), AT-AC (0.08%),

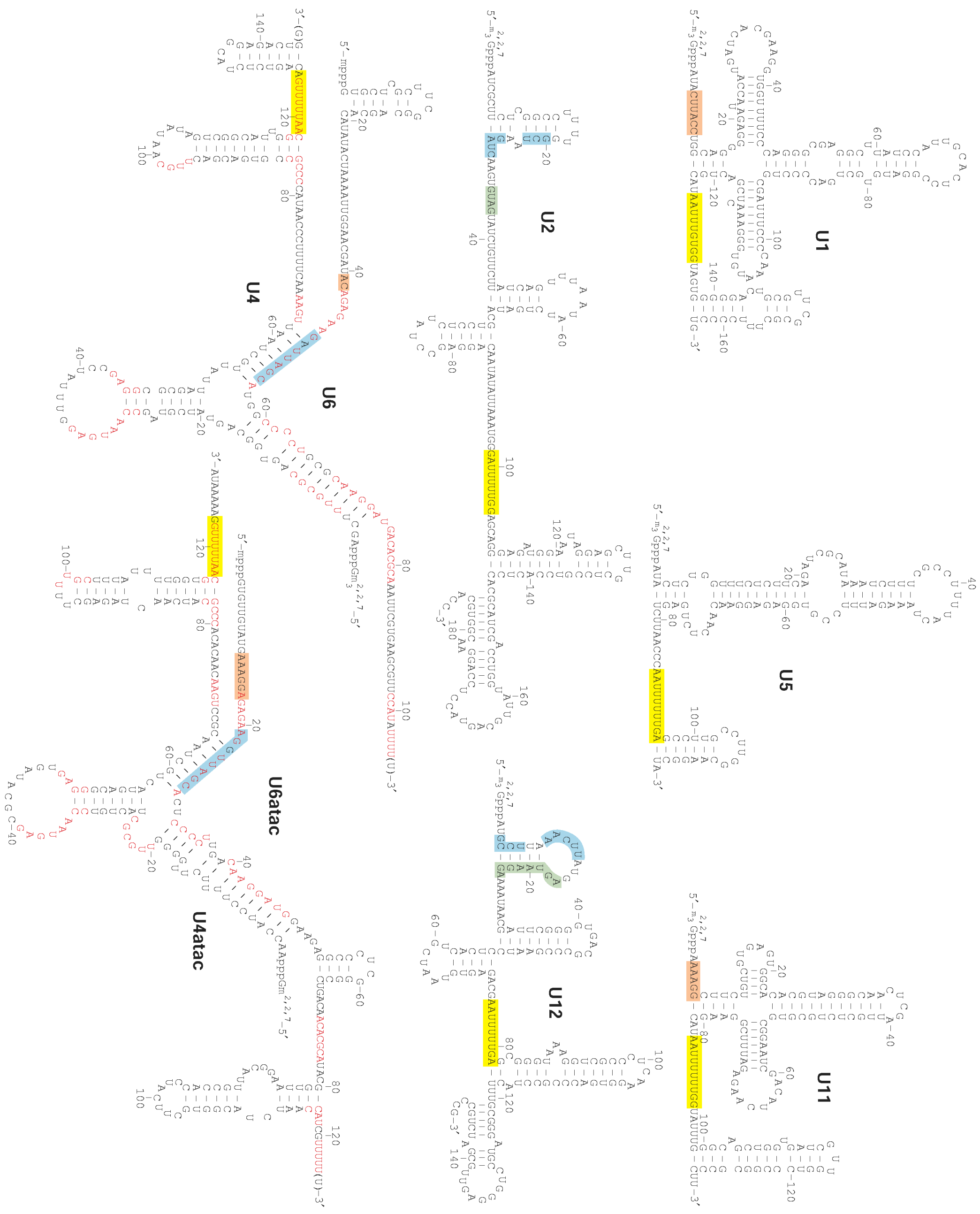


Figure 4.2: **Sequences and predicted secondary structures of the human spliceosomal snRNAs.** Similarities in secondary structure are apparent between the major- and the minor-class snRNA counterparts (U1 and U11, U2 and U12, and U4-U6 and U4ATAC-U6ATAC), despite substantial sequence divergence. The Sm-binding sites are shaded in yellow. Coloured boxes indicate sequences that are predicted to be involved in intermolecular RNA-RNA base-pairing interactions: 5′ splice site (orange), branch site (green), and U2-U6 or U12-U6ATAC helix I interactions (blue). Sequences in red represent stretches of four or more identical nucleotides between U4-U6 and U4ATAC-U6ATAC. Adapted from Patel and Steitz [2003].

other non-canonical (0.03%) and errors (0.06%). A more recent analysis of such frequencies for human, mouse and rat splice sites can be found in Table 4.3 on page 132 (Table 2 on page 114 of Abril *et al.* 2005). Biochemical studies showed that mutation of AT-AC to GT-AG termini did not interfere with splicing by the U12-dependent pathway. Instead, U12-dependent splicing is determined by the longer and more tightly constrained consensus sequences at the 5′ splice site and branch site of minor-class introns, as well as by the lack of a polypyrimidine tract upstream of the 3′ splice site [Dietrich *et al.*, 1997; Sharp and Burge, 1997; Burge *et al.*, 1998]. Therefore, the more suitable 'U12-type' nomenclature was adopted for this new class of introns.

4.1.2 The splicing process

For major-class introns, spliceosome assembly (see left pathway of Figure 4.3) is thought to begin with the association of the U1 and U2 snRNPs by base-pairing interactions with conserved sequences at the 5′ splice site (5′ss) and intron branch site, respectively [Reed, 1996]. During the spliceosome assembly, the U1 snRNP binds to the 5′ss via base base pairing between the splice site and the U1 snRNA. The 3′ splice site (3′ss) elements are bound by a special set of protein factors, SF1 (a branch-point binding protein, also called BBP in yeast), SF3, and a dimeric U2 snRNP auxiliary factor (U2AF). The 65 kDa subunit of U2AF binds to the polypyrimidine track. In at least some cases, the 35 kDa subunit of U2AF binds to the AG at the intron/exon junction. In mammalian cells, selection of the branch point is based primarily upon relative position, in the vast majority of cases the RNA branch forms 18–38 nucleotides upstream of the 3′ss. It is probable that this distance constraint reflects the requirement for the U2AF protein. The earliest defined complex in spliceosome assembly, called the commitment complex or E-complex (early), contains U1 and U2AF bound at the two intron ends [Burge *et al.*, 1999]. The E-complex is joined by the U2 snRNP, whose snRNA base-pairs at the branch point, to form the A complex. The U2 branch-site duplex protrudes outwards the adenosine residue, the 2′ hydroxyl group of which participates in the first nucleophilic atack.

The tri-snRNP complex of U5 and the base-paired U4-U6 then stably joins the prespliceosome [Konarska and Sharp, 1987] to form the B-complex, although there is evidence to suggest that U5 interacts upstream of the 5′ss at a much earlier stage [Wyatt *et al.*, 1992]. The B-complex undergoes a complicated rearrangement to form the activated spliceosome (B*-complex). This rearrangement is promoted by ATP-hydrolyzing protein factors that juxtapose the 5′ and 3′ss and form the catalytic core. U4-U6 duplexes unwind [Lamond *et al.*, 1988], and the U4 and U1 snRNPs are displaced, which allows U6 to form base-pairing interactions with the 5′ss [Wassarman and Steitz, 1992] and with a region of U2 that is near the U2 branch-site duplex [Datta and Weiner, 1991; Hausner *et al.*, 1990; Madhani and Guthrie, 1992; Wu and Manley, 1991]. The activated spliceosome catalyzes the first trans-esterification step of splicing and the C-complex is formed. The U5 snRNP has been shown to base-pair with sequences in both the 5′ and 3′ exons (see Figure 4.4). U5 is also believed to position the ends of the two exons for the second step of splicing [Wyatt *et al.*, 1992; Wassarman and Steitz, 1992; Newman and Norman, 1991, 1992; Sontheimer and Steitz, 1993]. After the second step has been completed, the ligated exons and a lariat intron are released, and the spliceosomal components dissociate and are recycled for further rounds of splicing.

Two general properties of the spliceosome are remarkable. First, it is conserved from

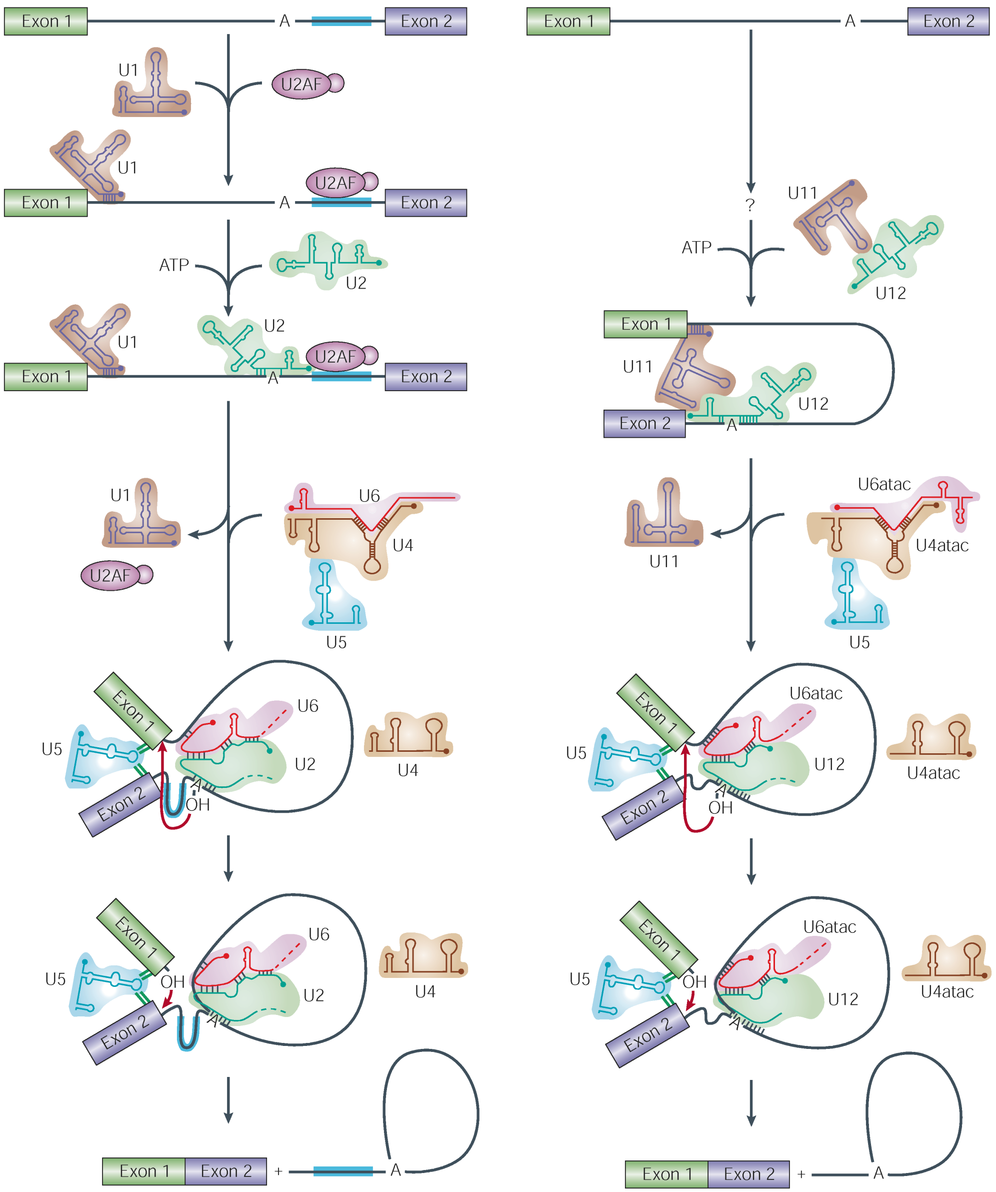


Figure 4.3: **Pathways of assembly and catalysis of U2 and U12 spliceosomes.** The major-class (*left*) and the minor-class (*right*) splicing pathways are shown side by side, highlighting their similarities and differences. The two pathways are mechanistically very similar. The primary differences occur during the early steps of spliceosome formation. The two trans-esterification reactions are indicated by red arrows. Each schematic snRNP is shown as a small nuclear RNA (not drawn to scale, with the 5′ terminus denoted by a dot) with the surrounding shaded area representing proteins. The polypyrimidine track of the major-class intron is shaded blue. Green bars represent interactions between the conserved loop of U5 and exon termini. Adapted from Patel and Steitz [2003].

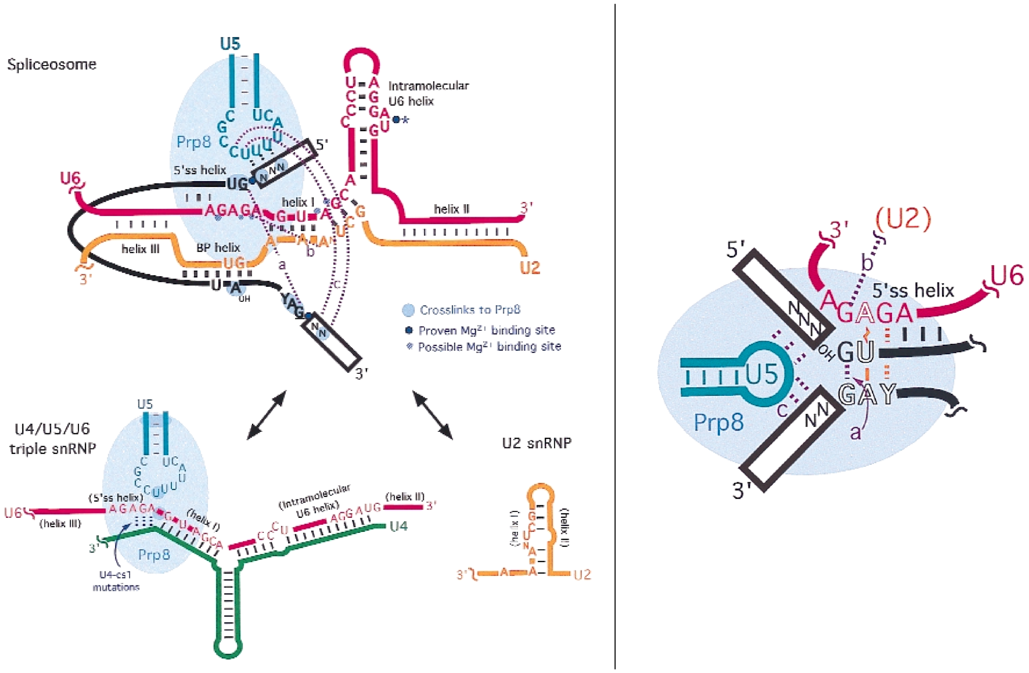


Figure 4.4: **Working model of RNA and Prp8 interactions in the catalytic core.** *Left panel*) Mutually exclusive interactions of U6 and U2 snRNAs in pre-assembled snRNPs are also shown. Large letters denote RNA sequences that are absolutely conserved in major, minor and trans spliceosomes from mammals, worms, plants, yeast and trypanosomes. Black lines denote Watson-Crick base-pairing interactions (the thinner lines denote interactions that are not absolutely conserved in all systems). Exons are drawn as rectangles, while the intron is depicted as a black line. *Right panel*) Some of the interactions of the active spliceosome are drawn for the second trans-esterification step of splicing: the 5′ splice site helix formed between U6 and the intron, and the interactions of the U5 conserved loop with exons. Purple dotted lines indicate tertiary interactions a, b, and c in both panels. Adapted from Collins and Guthrie [2000].

yeast to humans, both in its protein make-up and in its small nuclear RNAs (snRNAs), which have short, almost universally conserved sequences that are known to be juxtaposed to the reaction center during catalysis. Second, it is extraordinarily flexible, as it can excise introns of many different lengths and many different sequences. It is also subject to regulation, giving rise to alternatively spliced products in different cells or at different stages of development [Patel and Steitz, 2003].

The mechanism of U12-type splicing has been characterized *in vitro* [Tarn and Steitz, 1996]. Psoralen crosslinking studies provided evidence that U12 indeed forms a duplex with the minor-class branch site, apparently bulging the branch-point adenosine [Tarn and Steitz, 1996], which can reside at two different positions within the consensus site [McConnell *et al.*, 2002]. The minor-class splicing reaction proceeds through the same two-step pathway as the major reaction, which involves formation of a lariat intermediate [Tarn and Steitz, 1996]. Native gel electrophoresis of spliceosomal complexes allowed the initial characterization of the assembly pathway, which is shown in the right panel of Figure 4.3, and indicated that U11, U12 and U5 were components of the minor-class spliceosome [Tarn and Steitz, 1996]. Interaction of U11 with the 5′ splice site was later confirmed by site-specific crosslinking [Yu and Steitz, 1997].

U4ATAC and U6ATAC are two low-abundance snRNPs with copy numbers similar to those of U11 and U12. Although their sequences diverge significantly from those of U4 and U6 (see Figure 4.2), they predict analogous secondary structures and interactions with the pre-mRNA and other snRNAs. Crosslinking studies confirmed the predicted interactions between U4ATAC and U6ATAC [Yu and Steitz, 1997], between U6ATAC and the minor-class 5′ss [Tarn and Steitz, 1996], and between U6ATAC and U12 [Yu and Steitz, 1997]. This showed that the two spliceosomes undergo comparable dynamic rearrangements in which the snRNAs assume equivalent architectures, as shown in Figure 4.3.

In vivo evidence of the requirement of U12 minor-class splicing came from genetic suppression experiments, in which the deficient splicing of a minor-class intron containing two point mutations at the branch site was rescued by co-expression of a U12 snRNA with compensatory mutations [Hall and Padgett, 1996]. Similar genetic suppression experiments provided evidence for the *in vivo* interaction between the minor-class 5′ss with U11 [Kolossova and Padgett, 1997], and U6ATAC [Incorvaia and Padgett, 1998]. Fruit-flies that are homozygous for disruptions in U12 or U6ATAC genes do not survive early development, which indicates that the minor-class spliceosome is essential for organisms that harbour U12-type introns [Otake *et al.*, 2002]. Indeed, the presence of U12-type introns within most metazoan genomes indicates that an active U12-type splicing system is indispensable for the cells of most multicellular organisms [Patel and Steitz, 2003].

Of the snRNAs employed in splicing, only U5 snRNA is shared between the two spliceosomes, whereas the vast majority of the spliceosomal proteins appear to be shared [Will *et al.*, 1999, 2001; Schneider *et al.*, 2002; Luo *et al.*, 1999]. The U5 snRNP is unique in serving as a component of both spliceosomes, which indicates that it does not base-pair with sequences that differ between the two intron types. Although its role in the major-class spliceosome can involve base-pairing [Wyatt *et al.*, 1992; Wassarman and Steitz, 1992; Newman and Norman, 1991, 1992; Sontheimer and Steitz, 1993], proteins are known to support the juxtaposition of exons for the second step of splicing. Recent evidence that the protein components of U5 undergo marked remodeling during spliceosome activation [Makarov *et al.*, 2002] indicates that U5 has a pivotal role in recruiting common protein factors to the two spliceosomes.

4.1.3 Integrating splicing in the protein synthesis pathway

Throughout their lifetimes mRNAs exists, *in vivo*, as mRNA-protein particles (mRNPs). The associated proteins control every aspect of mRNA metabolism, from subcellular transport to translational efficiency to their rate of decay. Exactly which proteins associate with a particular mRNA depends on its sequence, its subcellular localization and its synthetic history. Furthermore, the complement of mRNA proteins evolves as the mRNA moves to different locations and is acted on by such processes as nuclear export and translation [Reichert *et al.*, 2002].

On the other hand, many pre-mRNA processing events—including 5′ end capping, splicing exons together, and 3′ end maturation by cleavage or polyadenylation—occur while the nascent RNA chain is being synthesized by RNA polymerase II. The RNApolII

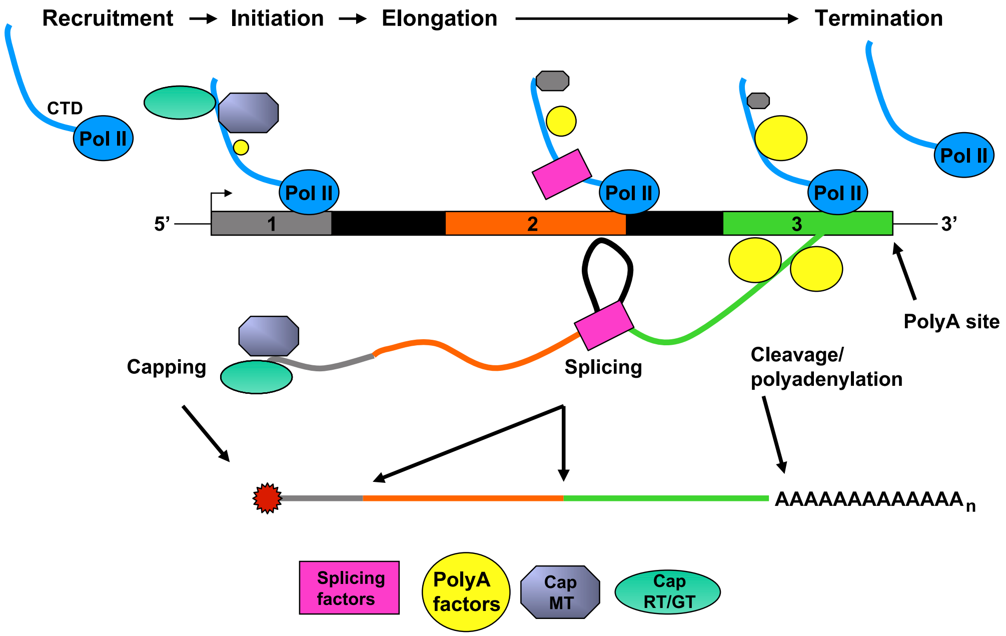


Figure 4.5: **The mRNA factory model.** Schematic representation of co-transcriptional processing. Processing factors interact with the RNApolII machinery via the carboxyl-terminal domain (CTD) of the largest subunit of RNApolII, Rpb1. The size of the symbols for processing factors corresponds to their levels of *in vivo* formaldehyde cross-linking, measured by ChIP experiments. Capping enzymes, RT, GT and MT, and 3′ end modifying factors (polyA related) are recruited at the 5′ ends of genes. As RNApolII traverses the gene, splicing factors associate with the transcription complex. Phosporylation of the Ser_2 and Ser_5 residues in the CTD heptad repeats decrease as the RNApolII advances. Exon numbers are marked in colored boxes, while introns are shown in black boxes. The red star represents the cap structure. Adapted from Zorio and Bentley [2004].

large subunit is equipped with a unique protein domain to tackle the job of directing co-transcriptional processing. This C-terminal domain (CTD) is composed of tandem repeats of the consensus heptad $Y_1S_2P_3T_4S_5P_6S_7$, which is conserved from fungi to humans [Corden and Ingles, 1992]. Deletion of the CTD in vertebrate cells reduces the overall level of transcription without necessarily affecting the accuracy of initiation. Deletion of the CTD inhibits all three major pre-mRNA processing steps in vertebrate cells: capping, splicing, and polyA site cleavage [McCracken *et al.*, 1997b,a]. The CTD functions as a landing pad for reversible interactions with RNA processing factors [Greenleaf, 1993] that serve to localize those factors close to their substrate RNAs and to act as a conduits for two-way communication with the polymerase.

As sketched in Figure 4.6, the cap binding complex (CBC) interacts with factors assembled on the 5′ss. Once the 3′ss has emerged from the elongating RNApolII, cross-intron interactions can be seen. U1 snRNP components, the U1-70K protein and Prp40/FBP11, can interact with SF1 and U2AF on the branch point, polypyrimidine track and 3′ss. Those interactions can be facilitated by protein-protein interactions mediated by serine/arginine-rich proteins (SR), which can act as exonic splicing enhancers. After that, two scenarios are possible: a new downstream 5′ss defining an internal exon or a downstream polyadenylation signal defining a terminal exon Goldstrohm *et al.* [2001].

Several examples of intronic and exonic *cis*-acting elements that are important for correct splice-site identification and that are distinct from the classical splicing signals have been described. These elements can act by stimulating (as do enhancers) or repressing (as do silencers) splicing, and they seem to be especially relevant for regulating alternative splicing [Cartegni *et al.*, 2002]. Exonic splicing enhancers (ESEs), in particular, appear to be very prevalent, and might be present in most, if not all, exons, including constitutive ones [Liu *et al.*, 1998; Schaal and Maniatis, 1999]. The analysis of the distribution of exonic splicing silencers (ESSs) revealed that ESSs appear more frequently in skipped exons, as well as in alternative 5′ and 3′ exons, in comparison with constitutive exons [Zhang and Chasin, 2004; Wang *et al.*, 2004]. In addition to ESEs and ESSs, intronic splicing enhancers (ISEs) and silencers (ISSs) are also an important part of the regulatory program in many alternative splicing events [Black, 2003]. ISEs and ISSs may also contribute to the definition of constitutive exons. In the human genome, RNA binding proteins are almost as abundant as transcription factors and the majority of them are of unknown function. Assignment of individual ESEs and ESSs to specific mediators will be essential for deciphering regulatory networks. Together with the rules for potential co-variation of ESEs and ESSs in exons, and by integrating the information with gene expression profiles, a true splicing regulatory code might be possible [Fu, 2004].

The spliceosome is believed to undergo some level of assembly and disassembly each time an intron is removed, but exactly how spliceosome recycling is achieved between successive introns in a given transcript remains a major unanswered question. It is not known whether a spliceosome is completely released from the transcription complex after two exons are ligated or whether some components remain associated with RNApolII and reused at downstream splice sites. Because the 5′ and 3′ splice sites are often quite distant from one another, splicing is the only processing event for which the RNA recognition sites are synthesized at different times. RNApolII elongates transcripts in a highly nonuniform way, punctuated by frequent pauses but with an average rate of $1 \sim 2$ kb/min [Conaway *et al.*, 2000]. This means that the 3′ss of a 30 kb intron would therefore be synthesized $15 \sim 30$ minutes after the 5′ss, time enough for this to bind the U1 snRNP and get ready for splicing. A 5′ss may pair with the first 3′ss to appear as proposed by the “first come first served” model [Aebi *et al.*, 1987]. Slow transcription would favor a proximal 3′ss over a distal site that only appears after a significant delay. Results, from tests on yeast and mammalian cells using RNApolII mutants and an inhibitor that slows down elongation [Howe *et al.*, 2003; de la Mata *et al.*, 2003], show that polymerases shifted the balance in favor of proximal over distal alternative 3′ss thereby reducing exon skipping. These results strongly support the idea that the effect of elongation rate on the lag time between the appearances of different splice sites can modulate alternative splicing. These experiments, therefore, argue for kinetic coupling of transcription and splicing. The effect of elongation rate on alternative splicing may explain how different promoter sequences can alter alternative splice site choices [Cramer *et al.*, 1997] since transcription factors bound to a promoter can influence the efficiency of elongation [Yankulov *et al.*, 1994].

Finally, nonsense-mediated mRNA decay (NMD) is an mRNA surveillance mechanism that has been described in organisms ranging from yeast to humans and ensures mRNA

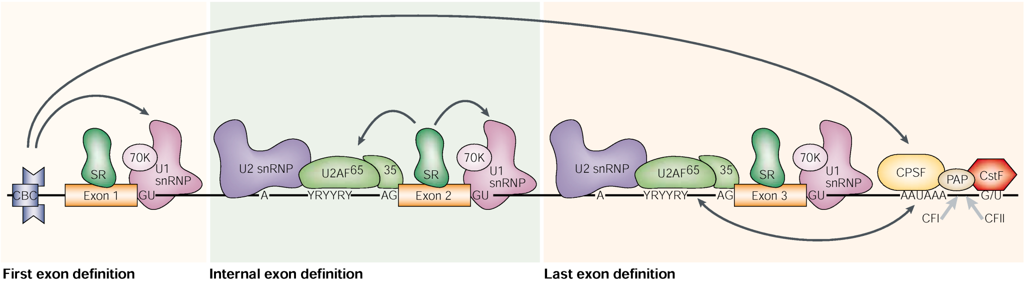


Figure 4.6: **Exon definition model in vertebrates.** Typically, exons are much shorter than introns in vertebrates. According to the exon-definition model, before introns are recognized and spliced-out, each exon is initially recognized by the protein factors that form a bridge across it. In this way, each exon, together with its flanking sequences, forms a molecular recognition module (arrows indicate molecular interactions). Adapted from Zhang [2002]. CBC, cap-binding protein; CFI/II, cleavage factor I/II; CPSF, cleavage and polyadenylation specificity factor; CstF, cleavage stimulation factor; PAP, poly(A) polymerase.

quality by selectively targeting mRNAs that harbour premature termination codons (PTCs) for rapid degradation [Hentze and Kulozik, 1999; Maquat, 1995, 2000]. PTCs that are introduced as a consequence of DNA rearrangements, frame shifts or nonsense mutations, or are caused by errors during transcription or splicing, can lead to non-functional or deleterious proteins. PTCs in higher eukaryotes are only recognized as such when they occur upstream of a boundary on the spliced mRNA that is situated approximately 55 nucleotides upstream of the last exon-exon junction [Maquat, 2000]. The prevalent view of the NMD mechanism is that the splicing process leaves a mark about 20 nucleotides upstream of each exon-exon boundary, in the form of an exon-junction complex (EJC), which in turn provides an anchor for up-frameshift suppressor proteins [Maquat, 2000; Hir *et al.*, 2000]. EJCs are formed by splicing-specific mRNP proteins, which associate with spliced mRNAs in a sequence-independent manner at a fixed distance upstream of exon-exon junctions [Hir *et al.*, 2000]. During the first round of translation, also known as pioneer round, of a normal mRNA, the stop codon is located downstream of the last mark, and all EJCs are displaced by elongating ribosomes [Ishigaki *et al.*, 2001]. During subsequent rounds of translation, the cap-binding complex is replaced by the eukaryotic initiation factor 4E (eIF4E) and the poly(A)-binding protein II (PABPII) is replaced by PABPI. New ribosomes no longer encounter EJCs and the mRNA is immune to NMD. However, when a PTC is present, ribosomes stop and fail to displace the downstream EJCs from the transcript. Then, interactions between the marking factors and components of the post-termination complex trigger mRNA decay. Moreover, intron containing genes are generally expressed at a significantly higher level in human cells than the same genes lacking introns [Buchman and Berg, 1988; Ryu and Mertz, 1989; Lu and Cullen, 2003; Nott *et al.*, 2003]. There is evidence that EJCs may be also responsible for the positive effect of splicing on gene expression [Wiegand *et al.*, 2003].

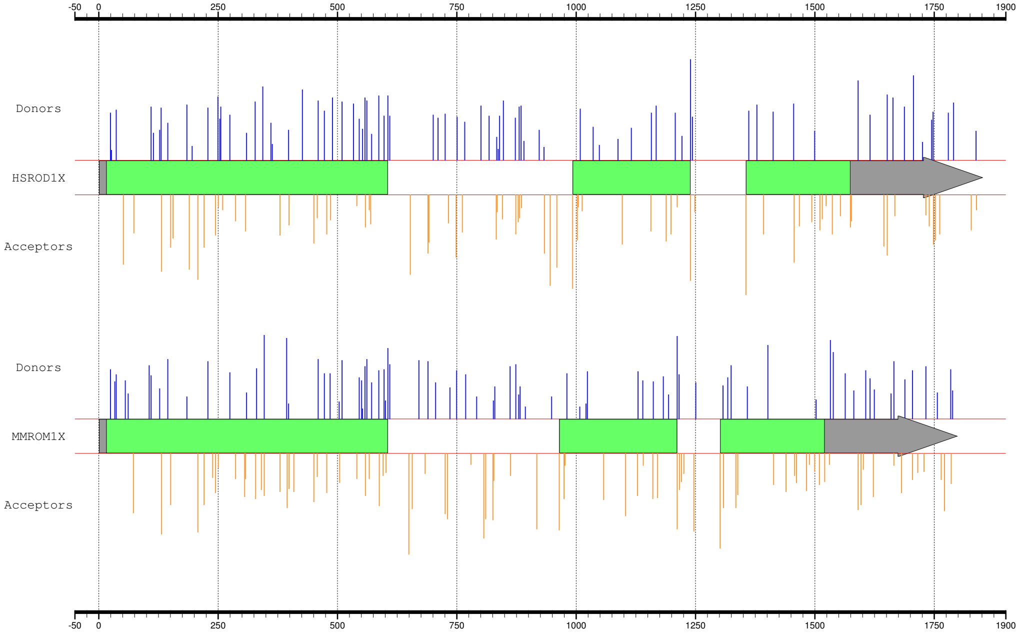


Figure 4.7: **Conservation of gene structure between human and mouse.** Human rod outer segment membrane protein 1 (Rom1, GenBank locus `HUMROD1X`) exonic structure is plotted on top, the orthologous gene structure in mouse (GenBank locus `MUSROM1X`) is shown below. Both genes have three coding exons. Exon and intron lengths are quite similar. A position-specific scoring matrix was used to calculate all potential splice sites along the sequence. Donors are shown as blue spikes and acceptors as orange ones, where the height of each spike represents the score for the corresponding site. A similar sites distribution is observed when comparing both genes. Although real splice sites have good scores, they are often not better than the surrounding predicted signals.

4.1.4 The conservation of exonic structure

Numerous regions that are conserved between human and mouse are found in introns [Hardison *et al.*, 1997]. Comparison of human chromosome 21 and the corresponding genomic sequences in mouse revealed that only one-third of the conserved blocks are exons, the other two-thirds being intronic and intergenic sequences [Dermitzakis *et al.*, 2002]. Hare and Palumbi [2003] describe that moderate rates of substitution rate heterogeneity, expected to result in part from mutational processes, can explain much of the conserved sequence observed in pairwise and three taxon comparisons, under a strictly neutral model of sequence evolution without indels. As a result, blocks of non-coding sequence conserved over long divergence times do not necessarily indicate selective constraints, even when observed across more than two taxa. However, they have found that half of the intron conservation observed cannot be explained by the typical levels of substitution rate heterogeneity in non-coding sequences. This strongly suggested that intronic sequences can play a larger functional role than previously realized.

After multiple complete sequences of eukaryotic genomes became available, comparative analyses revealed numerous introns that occupy the same position in orthologous genes from distant species [Fedorov *et al.*, 2002; Rogozin and Pavlov, 2003]. The great majority (>90%) of intron positions that are shared by phylogenetically distant eukaryotes—for example plants, fungi and metazoans—seem to reflect *bona fide* evolutionary conservation [Sverdlov *et al.*, 2005]. This is supported, for instance, by the observed dramatic differences between intron distributions in animal genomes. Those differences depend on non-local features of gene organization, such as the avoidance of short exons and the non-uniform distribution of introns accross the length of genes, for example preferential location of introns in the 5′ portions of genes in many species [Smith, 1988; Stoltzfus *et al.*, 1997; Mourier and Jeffares, 2003; Sverdlov *et al.*, 2004]. Therefore, it seems unlikely that those features had a substantial impact on the long-term evolution of introns [Sverdlov *et al.*, 2005].

Recent large scale comparative analyses have reported extraordinary conservation of the exonic structure between human and mouse orthologous genes [Roy, 2003]. Almost all of the protein-coding genes (99%) in human align with homologs in mouse, and over 80% are clear 1:1 orthologs. In most cases, the intron-exon structures are highly conserved [Waterston *et al.*, 2002], as can be seen, for instance, in Figure 4.7. Estimates of the proportion of 1:1 orthologs between mouse and rat lie between 86 and 94%. Surprisingly, a similar proportion, 89 to 90% of rat genes possessed a single orthologue in the human genome [Gibbs *et al.*, 2004]. About 60% of the chicken protein-coding genes have a single human orthologue [Hillier *et al.*, 2004]. Furthermore, the extent of conservation of alternative splicing between human and mouse is high. It has been suggested that patterns of alternative splicing are conserved at similar levels to genes and gene structures, with overall conservation estimates of 61% of alternative and 74% of constitutive splice junctions [Thanaraj *et al.*, 2003]. Sorek and Ast [2003] have reported that 77% of the conserved alternative spliced exons between human and mouse were flanked on both sides by long conserved intronic sequences. In comparison, only 17% of the conserved constitutively spliced exons were flanked by such conserved sequences. These results suggest that the function of many of the intronic sequence blocks that are conserved between human and mouse is the regulation of alternative splicing [Arian Smith, *pers. communic.*].

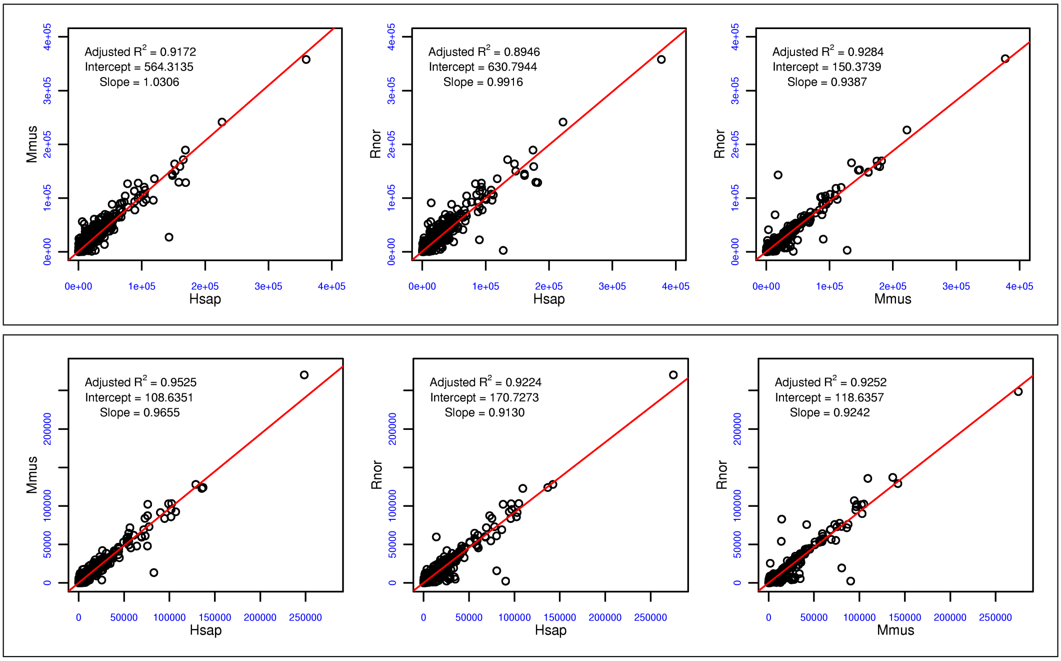


Figure 4.8: **Human/mouse/rat scatterplots for orthologous GT-AG intron lengths.** Upper panels show pair-wise comparisons of orthologous intron lengths. Repeat lengths have been removed from the corresponding total intron lengths in the pair-wise comparison in the lower panels ($N = 6,261$ orthologous introns).

4.2 The Comparative Analysis of Mammalian Gene Structures

Preliminary comparative analyses of the human and mouse gene structures for a set of 1,506 pais of orthologous genes are shown in the section entitled "Conservation of gene structure" on page 43 (page 551 of Waterston *et al.* 2002). In what follows, we describe our major contributions to the understanding of the exonic structures and the splice signals of the orthologous genes of human, mouse and rat.

4.2.1 Intron length and repeats

Of a set of 6261 human/mouse/rat orthologous introns we have computed the average intron length for each species. Results are shown in table 4.1. On average, introns are longer in human than in rodent and rat introns appear to be longer than those of mouse. Our numbers for human and mouse intron lengths are comparable to those reported in Waterston *et al.* [2002]. There is strong correlation, however, between the length of orthologous introns in different species (the correlation coefficient is about 0.90 between human and rodent, and 0.94 between mouse and rat. The correlation coefficient between length of orthologous exons is in all cases larger than 0.99).

	Intron Length		**Percentage of Intron Length**		
Species	with repeats	without repeats	in all repeats	in ancient repeats	in other repeats
human	4,765	2,747	42.57	15.70	26.87
mouse	3,770	2,558	32.60	4.72	27.88
rat	4,102	2,872	30.38	4.63	26.69

Table 4.1: **Intron length and proportion of repetitive DNA in mammalian introns.**

Differences in length between human and mouse orthologous introns are attributable to a larger fraction of repetitive DNA in human than in rodent introns: while DNA in repeats accounts for 43% of the human intron sequences, it accounts for only around 30% in rodent introns (see table 4.1). Therefore, when subtracting the number of bases masked by the program `RepeatMasker` [see page 215, on Web Glossary; Smit *et al.*, 1996–2004] differences in length between human and rodents reduce notably (see table 4.1), with rat introns having the highest proportion of non-repetitive DNA.

Since it may be argued that the orthologous intron dataset is a too small and biased sample of all introns in these organisms, we have computed intron length for all genes in the REFSEQ collection [Pruitt and Maglott, 2001; Pruitt *et al.*, 2005], before applying the filtering protocol; see the corresponding methods section on page 135 (page 117 of Abril *et al.* 2005). Average intron lengths, including and excluding masked nucleotides are, respectively, 5,632 and 3,247 in human (177,931 introns), 4,423 and 2,996 in mouse (104,591 introns), and 4,933 and 3,451 in rat (37,043 introns). Therefore, even though our data set of orthologous introns appears to be biased towards shorter introns, the bias is similar in all organisms and does not affect the fraction of intronic DNA in repeats.

Interestingly, longer human introns do not appear to be the result of repeat expansion in the human lineage, but rather of the selective loss of ancient repeats in rodents. We have computed the fraction of intron sequence in repetitive DNA separately for ancient and recent repeats. As can be seen in table 4.1, the fraction of intronic DNA in recent repeats is essentially identical in the three species, suggesting that the dynamics of new repeat generation have not changed after the divergence of the lineages leading to rodent and human. However, ancient repeats are much more abundant in human introns (16% of the sequence) than in rodent introns (5% of the sequence), indicating that repeated sequences are eliminated much faster in rodents than in human. Although repeats appear to be generated slightly faster and to be lost slightly slower in the rat than in the mouse genome, repeat abundance does not account for the notable difference in intron length observed between these two rodent species. We have to take into account that due to a higher substitution level in the rodent lineage, `RepeatMasker` results can be biased to find human ancient repeats. At any rate, the youngest ancient repeats in mouse and rat have a 35–40% substitution level, which is on the border of what `RepeatMasker` can detect, while in the human genome these repeats have about 15% substitution and are reocgnized very easily [Waterston *et al.*, 2002]. `BLASTN` results from cross-matching the repeats found in all the orthologous introns against the intronic sequences of each other species, were supporting our hypothesis.

We did not continue the analyses reported in this section as a broader analysis of repeats, whole genome based, was presented for the mouse and the rat genomes [Waterston *et al.*, 2002; Gibbs *et al.*, 2004]. The large scale deletion level of non-essential DNA in rodents was much larger than in the human lineage. This results further in a reduced number of ancient repeats in the current rodent genomes; for instance, approximately 50% of the ancestral junk DNA, as it was at the human-mouse split, has been lost in mouse and only about 25% in human.

4.2.2 Sequence conservation at orthologous splice sites

See section entitled "Conservation of intronic splice signals" on page 119 (page 505 of Gibbs *et al*. 2004).

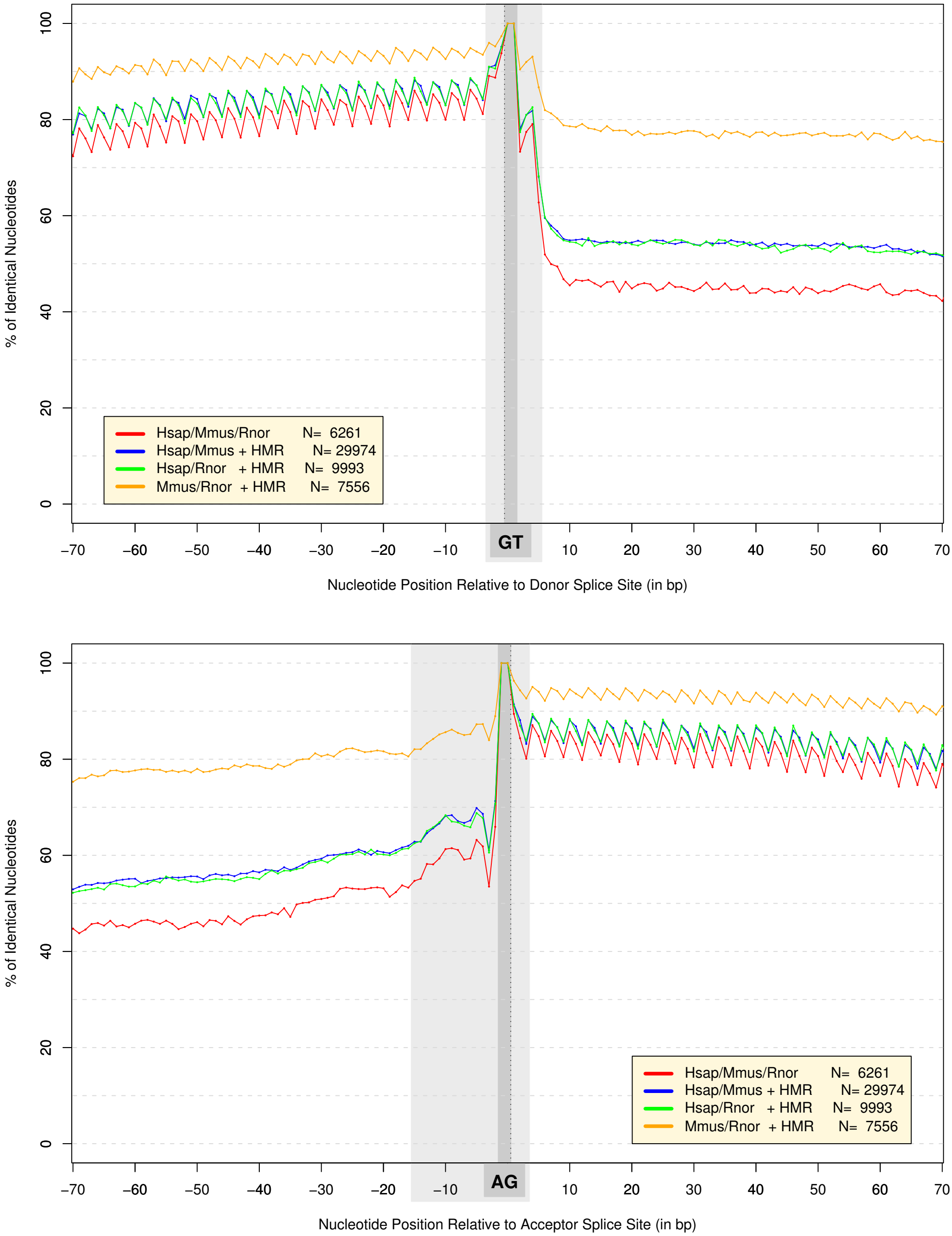


Figure 4.9: **Human/mouse/rat sequence conservation at orthologous GT-AG splice sites.** Sequence conservation for donor sites [supplementary materials Figure 8 of Gibbs *et al*. 2004] and acceptor sites [supplementary materials Figure 7 of Gibbs *et al*. 2004] are shown in upper and lower panels respectively.

4.2.3 RGSPC, *Nature*, 428(6982):493–521, 2004

PubMed Accession:

http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&list_uids=15057822&dopt=Abstract

Journal Abstract:

http://www.nature.com/cgi-taf/DynaPage.taf?file=/nature/journal/v428/n6982/abs/nature02426_fs.html

Supplementary Materials:

See Section 4.3.2 and the following URL:
http://www.nature.com/nature/journal/v428/n6982/suppinfo/nature02426.html

NOTE: Because of copyright restrictions, we cannot offer the article, please follow links for fulltext.

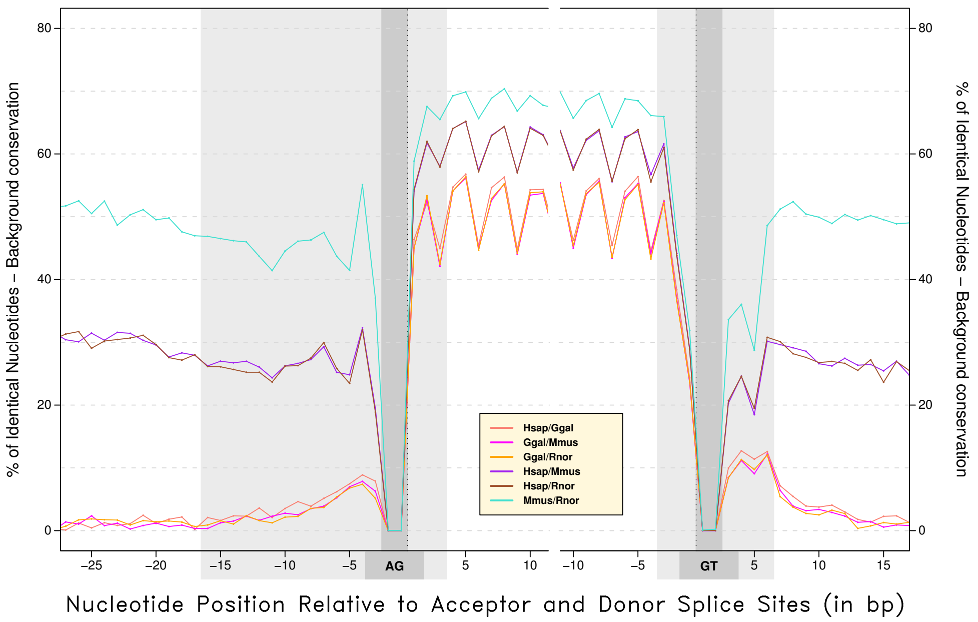


Figure 4.10: **Human/mouse/rat/chicken relative conservation over GT-AG splice site consensi.** The x-axis shows idealized base position from intron through exon to intron. The gray areas show the regions where expected conservation from the presence of splice site consensi was removed. Unlike inter-mammal comparisons, the chicken-mammal comparison shows a higher relative conservation rate at the splice sites than in the introns. Included as supplementary materials Figure 1 on Hillier *et al.* [2004].

4.3 The Comparative Analysis of Splice Sites in Vertebrates

4.3.1 Conservation of mammals and chicken orthologous splice sites

See section entitled "Evolutionary conservation of gene components" on page 142 (page 698 of Hillier *et al.* 2004).

Only the orthologous U12 introns of the four species were displayed in Figure 4.11. Further orthologous sets, including pair-wise and triads, are available at the supplementary materials web page (see page 213 on Web Glossary). It is worth to note that in the fourth example, the 16th intron of mouse gene *NM_007459* does not seem to be conforming to the U12 donor pattern. But it is not a case of conversion between U2 and U12 splice sites, just displacing the splice sites two nucleotides upstream we recover the U12 donor pattern and the overall alignment of the exonic regions improves.

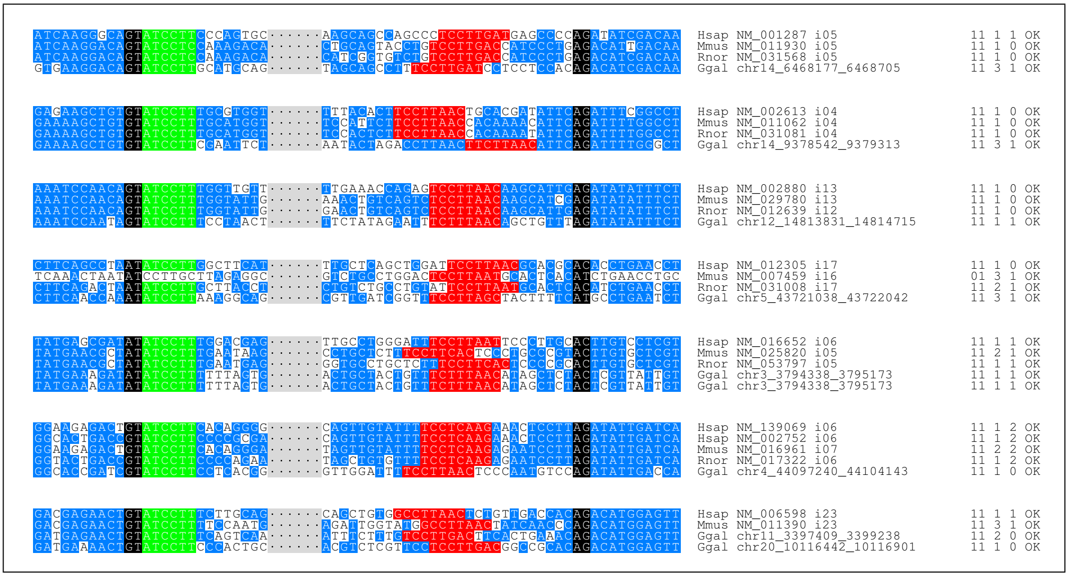

ATCAAGGGCAGTATCCTTCCCAGTGC······AAGCAGCCAGCCCTCCTTGATGAGCCCCAGATATCGACAA Hsap NM_001287 i05 11 1 1 OK
ATCAAGGACAGTATCCTCCAAAGACA······CTGCAGTACCTGTCCTTGACCATCCCTGAGACATTGACAA Mmus NM_011930 i05 11 1 0 OK
ATCAAGGACAGTATCCTCCAAAGACA······CATCGGTGTCTGTCCTTGACCATCCCTGAGACATCGACAA Rnor NM_031568 i05 11 1 0 OK
GTGAAGGACAGTATCCTTGCATGCAG······TAGCAGCCTTTCCTTGATCCTCCTCCACAGACATCGACAA Ggal chr14_6468177_6468705 11 3 1 OK
GAGAAGCTGTGTATCCTTTGCGTGGT······TTTACACTTCCTTAACTGCACGATATTCAGATTTCGGCCT Hsap NM_002613 i04 11 1 0 OK
GAAAAGCTGTGTATCCTTTGCATGGT······TCCATTCTTCCTTAACCACAAAACATTCAGATTTTGGCCT Mmus NM_011062 i04 11 1 0 OK
GAAAAGCTGTGTATCCTTTGCATGGT······TCCACTCTTCCTTAACCACAAAATATTCAGATTTTGGCCT Rnor NM_031081 i04 11 1 0 OK
GAAAAGCTGTGTATCCTTCGAATTCT······AATACTAGACCTTAACTTCTTAACATTCAGATTTTGGGCT Ggal chr14_9378542_9379313 11 3 1 OK
AAATCCAACAGTATCCTTTGGTTGTT······TTGAAACCAGAGTCCTTAACAAGCATTGAGATATATTTCT Hsap NM_002880 i13 11 1 0 OK
AAATCCAACAGTATCCTTTGGTATTG······AAACTGTCAGTCTCCTTAACAAGCATCGAGATATATTTCT Mmus NM_029780 i13 11 1 0 OK
AAATCCAACAGTATCCTTTGGTATTG······GAACTGTCAGTCTCCTTAACAAGCATTGAGATATATTTCT Rnor NM_012639 i12 11 1 0 OK
AAATCCAATAGTATCCTTTCCTAACT······TTCTATAGAATTTCTTTAACAGCTGTTTAGATATATTTCT Ggal chr12_14813831_14814715 11 1 1 OK
CTTCAGCCTAATATCCTTGGCTTCAT······TTGCTCAGCTGGATTCCTTAACGCACGCACACCTGAACCT Hsap NM_012305 i17 11 1 0 OK
TCAAACTAATATCCTTGCTTAGAGGC······GTCTGCCTGGACTCCTTAATGCACTCACATCTGAACCTGC Mmus NM_007459 i16 01 3 1 OK
CTTCAGACTAATATCCTTGCTTACCT······CTGTCTGCCTGTATTCCTTAATGCACTCACATCTGAACCT Rnor NM_031008 i17 11 2 1 OK
CTTCAACCAAATATCCTTAAAGGCAG······CGTTGATCGGTTTCCTTAGCTACTTTTCATGCCTGAATCT Ggal chr5_43721038_43722042 11 3 1 OK
TATGAGCGATATATCCTTTGGACGAG······TTGCCTGGGATTTCCTTAATTCCCTTGCACTTGTCCTCGT Hsap NM_016652 i06 11 1 1 OK
TATGAACGCTATATCCTTTGAATAAG······CCTGCTCTTTCCTTCACTCCCTGCCCGTACTTGTGCTCGT Mmus NM_025820 i05 11 2 1 OK
TATGAACGCTATATCCTTTGAATGAG······GGTGCCTGCTCTTTCCTTCACTCCCCGCACTTGTGCTCGT Rnor NM_053797 i05 11 1 1 OK
TATGAAAGATATATCCTTTTTTAGTG······ACTGCTACTGTTTCTTTAACATAGCTCTACTCGTTATTGT Ggal chr3_3794338_3795173 11 1 1 OK
TATGAAAGATATATCCTTTTTTAGTG······ACTGCTACTGTTTCTTTAACATAGCTCTACTCGTTATTGT Ggal chr3_3794338_3795173 11 1 1 OK
GGAAGAGACTGTATCCTTCACAGGGG······CAGTTGTATTTTCCTCAAGAAACTCCTTAGATATTGATCA Hsap NM_139069 i06 11 1 2 OK
GGCACTGACCGTATCCTTCCCCGCGA······CAGTTGTATTTTCCTCAAGAAACTCCTTAGATATTGATCA Hsap NM_002752 i06 11 1 2 OK
GGAAGAGACTGTATCCTTCACAGGGA······TAGTTGTATTTTCCTCAAGAGAATCCTTAGATATTGATCA Mmus NM_016961 i07 11 2 2 OK
GGTACTGACCGTATCCTTCCCCAGAA······TAGCTGTGTTTTCCTCAAGAGAATCCTTAGATATTGATCA Rnor NM_017322 i06 11 1 2 OK
GGCACCGATCGTATCCTTCCTCACGG······GTTGGATTTTCCTTAACTCCCAATGTCCAGATATTGACCA Ggal chr4_44097240_44104143 11 1 0 OK
GACGAGAACTGTATCCTTTCTTGCAG······CAGCTGTGGCCTTAACTCTGTTGACCACAGACATGGAGTT Hsap NM_006598 i23 11 1 1 OK
GACGAGAACTGTATCCTTTTCCAATG······AGATTGGTATGGCCTTAACTATCAACCCAGACATGGAGTT Mmus NM_011390 i23 11 3 1 OK
GATGAGAACTGTATCCTTTCAGTCAA······ATTTCTTTGTCCTTGACTTCACTGAAACAGACATGGAGTT Ggal chr11_3397409_3399238 11 2 0 OK
GATGAAAACTGTATCCTTCCCACTGC······ACGTCTCGTTCCTCCTTGACGGCCGCACAGACATGGAGTT Ggal chr20_10116442_10116901 11 1 0 OK

Figure 4.11: **Human, mouse, rat and chicken orthologous U12 intron sets.** Ungapped alignments of the donor (-10 to +16 around the 5′ splice sites) and the acceptor (-30 to +10 around the 3′ splice sites) sequences for all the orthologous U12 intron sets were drawn using TeXshade [Beitz, 2000]. Splice sites core signals are highlighted in a black box, the conserved U12 donor sequence (+3 to +8) is marked in green, sequence hits to the U12 branch point are colored in red, while conserved nucleotides at a given position are shown with a blue background.

4.3.2 Abril *et al*, *Genome Research*, 15(1):111–119, 2005

PubMed Accession:

`http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&list_uids=15590946&dopt=Abstract`

Journal Abstract:

`http://www.genome.org/cgi/content/abstract/15/1/111`

Supplementary Materials:

`http://genome.imim.es/datasets/hmrg2004/`

Chicken Special/Letter

Comparison of splice sites in mammals and chicken

Josep F. Abril, Robert Castelo, and Roderic Guigó[1]

Grup de Recerca en Informàtica Biomèdica, Institut Municipal d'Investigació Mèdica, Universitat Pompeu Fabra, and Programa de Bioinformàtica i Genòmica, Centre de Regulació Genòmica, C/ Dr. Aiguader 80, E-08003 Barcelona, Catalonia, Spain

We have carried out an initial analysis of the dynamics of the recent evolution of the splice-sites sequences on a large collection of human, rodent (mouse and rat), and chicken introns. Our results indicate that the sequences of splice sites are largely homogeneous within tetrapoda. We have also found that orthologous splice signals between human and rodents and within rodents are more conserved than unrelated splice sites, but the additional conservation can be explained mostly by background intron conservation. In contrast, additional conservation over background is detectable in orthologous mammalian and chicken splice sites. Our results also indicate that the U2 and U12 intron classes seem to have evolved independently since the split of mammals and birds; we have not been able to find a convincing case of interconversion between these two classes in our collections of orthologous introns. Similarly, we have not found a single case of switching between AT-AC and GT-AG subtypes within U12 introns, suggesting that this event has been a rare occurrence in recent evolutionary times. Switching between GT-AG and the noncanonical GC-AG U2 subtypes, on the contrary, does not appear to be unusual; in particular, T to C mutations appear to be relatively well tolerated in GT-AG introns with very strong donor sites.

[Supplemental material is available online at www.genome.org. The following individuals kindly provided reagents, samples, or unpublished information as indicated in the paper: P. Bork and I. Letunic.]

Protein-coding genes are characteristically interrupted by introns in the genome of higher eukaryotic organisms. While intron function and origin has been debated at length (de Souza 2003; Fedorova and Fedorov 2003; Roy et al. 2003), recent comparative analyses show an abundance of conserved elements in intronic sequences (for instance, see Dermitzakis et al. 2002; Hare and Palumbi 2003). This strongly suggests that introns are rich in elements playing functional, probably regulatory, roles (Mattick 2001). Splicing of introns is found in all main branches of eukaryotes, that is, animals, plants, fungi, and protozoa, indicating an early origin of splicing within eukaryotes, or the existence, in the pre-eukaryotic world, of a precursor of splicing. Indeed, the two major molecular mechanisms by means of which splicing is produced, U2- and U12-dependent, seem to have evolved independently prior to the divergence of the animal and plant kingdoms (Burge et al. 1998; Zhu and Brendel 2003).

Within each of these two classes of splicing, sequence features involved in intron specification are essentially conserved across eukaryotes. In both classes, the sequence information needed to specify the 5′ and 3′ splice sites—hereafter also described as donor and acceptor sites respectively—is largely confined to their surrounding region (see Fig. 1). Conserved sequences in these regions interact with the splicing machinery to promote the assembly of the spliceosome and activate the biochemical pathway that leads to the production of the spliced mRNA (for review, see Burge et al. 1999). Despite the strong conservation, the sequence of splicing signals does not carry enough information to unequivocally specify introns in the large sequence of the pre-mRNA transcripts, occasionally hundreds of thousands of nucleotides long; and recent research suggests that signals other than those in the region of the splice sites play a role in the definition of the intron boundaries (for review, see Caceres and Kornblihtt 2002; Cartegni et al. 2002; Black 2003).

Thus, in eukaryotic organisms, splicing introduces an additional level of decoding—prior to translation—on the sequence of the primary RNA transcript. There is a fundamental difference, however, between the genetic code—the mapping of nucleotide sequences (triplets) into 20 (or more) amino acids—and the splicing code—the mapping of nucleotide sequences into 3′ and 5′ intron boundaries. The genetic code is essentially deterministic; within a given species, a given triplet in the mRNA sequence results always in the same amino acid—the dual role in selenoproteins of the TGA triplet as stop and selenocysteine codon probably the most notable of all exceptions (for instance, see Kryukov et al. 2003). The splicing code, in contrast, is inherently stochastic; the probability of a splicing sequence in the primary transcript to participate in the definition of an intron boundary ranges from zero to one, and it is conditioned to very many different factors (which could be other sequences—maybe distant). The tissue-specific distribution of relative abundances of alternative splicing products (Xu et al. 2002; Yeo et al. 2004), for instance, reflects this nondeterministic nature of the splicing code.

The stochasticity of the splicing code offers opportunities for evolution that are absent in the highly deterministic genetic code. The availability of an increasing number of eukaryotic genomes makes it possible to investigate such an evolutionary process. Here, we report on findings obtained by comparing a large collection of orthologous introns (introns occurring at equivalent locations in orthologous genes) and their defining splice sites in human, mouse, rat, and chicken. Our results provide insights into the dynamics of the evolution of splice-site sequences during the most recent period of the history of life on earth.

Results

In this section, we first report results concerning interconversion between the two major classes of introns, U2 and U12, and subtype switching within each class. Then, we report on the com-

[1]**Corresponding author.**
E-mail rguigo@imim.es; fax 34-93-221-32-37.

Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.3108805. Article published online before print in December 2004.

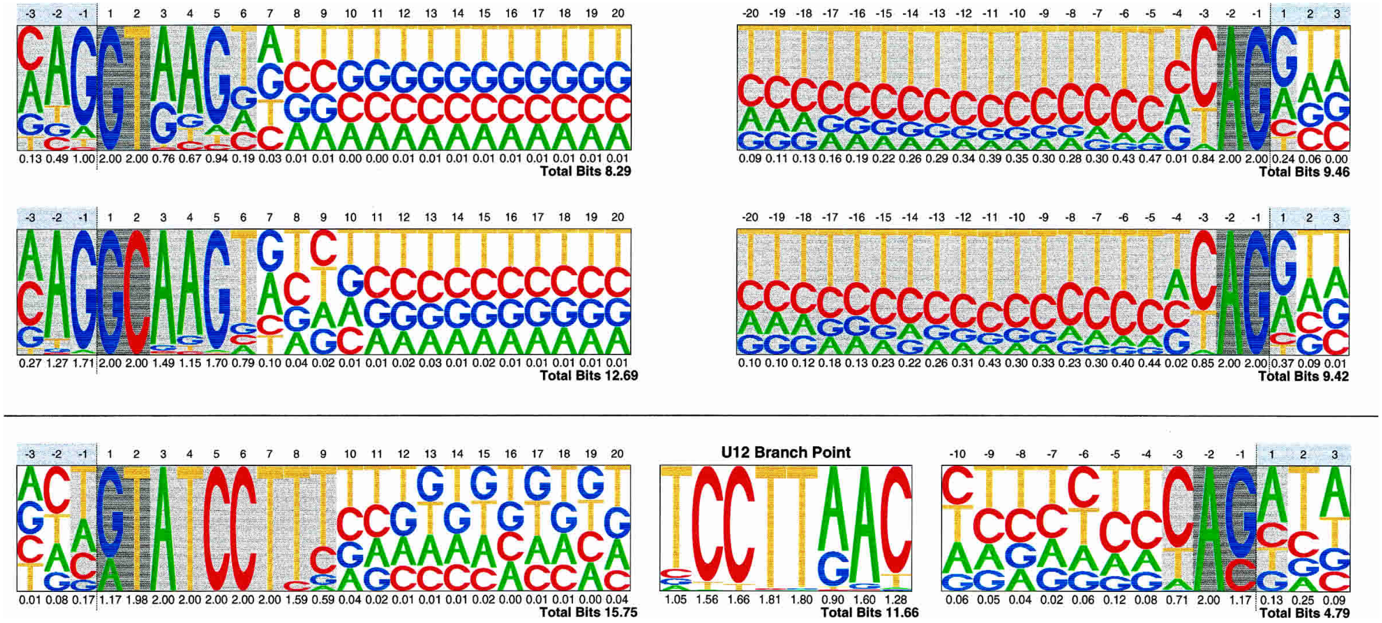


Figure 1. Donor and acceptor sites' pictograms. Pictograms of the donor (*left*) and the acceptor (*right*) site sequences for the U2 (*top*) and U12 (*bottom*) splice sites. The sequence plots for GT-AG and GC-AG U2 introns are given separately. The conserved sequence of the U12 branch point is also shown. From human, mouse, rat, and chicken RefSeq genes, a total number of 337,336, 2506, and 935 splice-site sequences from CDS introns from Ensembl were included in GT-AG, GC-AG, and U12 splice site sets, respectively, to produce the corresponding pictograms.

parison of splice-site sequences in human, rodents, and chicken. We have compared the overall sequence patterns of splice sites and investigated the level of sequence conservation between orthologous splice sites.

The analyses described here are very sensitive to the identification of true orthologous introns, as well as to the prediction of correct splice boundaries, particularly in the case of the noncanonical U12 introns. Because U12 introns constitute only a tiny fraction of all eukaryotic introns, computational gene prediction methods ignore them. Therefore, in absence of good cDNA coverage, computational gene catalogs are likely to heavily misrepresent them. Such is the case in the chicken genome. In an effort to conciliate the amount of data with reliability, we have resorted to different data sets to perform different types of analyses. Gene predictions from the RefSeq collection (Pruitt et al. 2003)—a collection of genes with good cDNA support—have been used for interspecific analysis of splice-site sequence patterns and for the identification and analysis of mammalian U12 introns. However, there are very few chicken genes in RefSeq. The larger—but strongly biased toward GT-AG canonical U2 introns—Ensembl collection (Birney et al. 2004; http://www.ensembl.org) has been used for interspecific comparison of splice-site patterns. A set of mammalian–avian curated orthologous introns—referred to as the HMRG set in this work (see Methods section)—has been used for the comparison of orthologous splice-site sequences. Table 1 describes the sizes of the data sets used in this study.

Intron classes

Two distinct types of pre-mRNA introns are found in most higher eukaryotic organisms (Sharp and Burge 1997). They differ in the spliceosome complex that excise them during RNA processing. More than 99% of eukaryotic introns are spliced by the U2 spliceosome, while a minor class are spliced by the U12 spliceosome. U2 and U12 introns differ in the conserved sequences flanking their splice sites (see Fig. 1). Vertebrate U2 introns are characterized by the highly variable consensus [CA]AG/GT[AG]AGT at the donor (5′) site, (where [CA] means C or A, and / denotes the exon–intron boundary) and by a polypyrimidine-rich stretch between the acceptor site and a poorly conserved branch point. The branch point and the acceptor site are usually separated by 11–40 nucleotides, although cases are known where they can be over 100 nucleotides apart (Helfman and Ricci 1989; Smith and Nadal-Ginard 1989). U2 introns almost always exhibit the conserved GT and AG dinucleotides at the 5′ and 3′ intron boundaries, respectively. The only remarkable exception is the existence of U2 GC-AG introns, which appears with a frequency <1% (Burset et al. 2001).

U12 introns are characterized by a strong consensus/[AG]TATCCTT at the donor site, and TCCTT[AG]AC at the branch point. They also lack the polypyrimidine tract upstream of the acceptor site, characteristic of U2 introns. Also, in contrast to U2 introns, the distance between this acceptor site and the branch point is consistently short, between 10 and 20 nucleotides (Dietrich et al. 2001). Although initially discovered because of the unusual AT and AC dinucleotides at the 3′ and 5′ splice sites (Jackson 1991; Hall and Padgett 1994), it was later shown that U12 introns can exhibit a variety of terminal dinucleotides, the vast majority, however, are GT-AG or AT-AC (Dietrich et al. 1997; Sharp and Burge 1997; Levine and Durbin 2001; Zhu and Brendel 2003). Subtype switching within U12 introns, as well as conversion from U12 to U2 introns, has been documented (Burge and Karlin 1998), although amazing stability has been reported for U12 introns over very large evolutionary times (Zhu and Brendel 2003).

We have used the U12 donor site and branch point patterns above to identify U12 introns in the human and rodent RefSeq collections (see Methods). Table 2 lists the resulting frequencies of the different splice classes, and subtypes within each class. Numbers are consistent with those previously published (Burset et al. 2001; Levine and Durbin 2001). Identification of U12 introns was not attempted in chicken because of the small size of the RefSeq database for this organism. Figure 1 uses sequence pictograms to display the consensus for GT-AG U2 splice signals in mammals and chicken. It also displays the mammalian consensus for GC-AG U2 and U12 splice sites. In sequence pictograms (Schneider and Stephens 1990; Burge et al. 1999) the frequencies of the four nucleotides at each position along the signal are represented by the heights of their corresponding letters. The information content (intuitively, the deviation from random composition) is computed at each position, and summed up along the signal. The larger the information content, the more conserved the signal.

Table 1. **Summary of initial data and filtered orthologs sets.**

(A) Initial data sets

	Ensembl[a]			UCSC genome browser[b] RefSeq[c]		
Species	**Version**	**Genes**	**Introns**	**Version**	**Genes**	**Introns**
human[d]	v19.34a	33,633	284,125	HGv16/NCBI34	21,744	206,814
mouse[e]	v19.30	30,665	218,163	MGSCv4/NCBI32	17,988	139,258
rat[f]	v19.3a	28,545	192,459	RGSCv3.1	4877	43,393
chicken[g]	v22.1.1	28,491	252,226	CGSCv2	1496	12,632

(B) Filtered orthologs

	Sets	**Genes**	**Introns**	
Total	human	6043	48,939	(out of 51,876)
	mouse	5680	45,543	(out of 47,193)
	rat	1847	13,929	(out of 14,245)
Orthologs	human/mouse	5550	44,119	
	human/rat	1737	13,259	
	mouse/rat	1416	9655	
Triads	human/mouse/rat	1283	8895	

(A) Initial data sets: the initial pool of genes/introns from which we filtered all the data sets for this work ([a]Birney et al. 2004; [b]Karolchik et al. 2003; [c]Pruitt et al. 2003; [d]Lander et al. 2001; [e]Waterston et al. 2002; [f]Rat Genome Sequencing Project Consortium 2004; [g]International Chicken Genome Sequencing Consortium 2004).
(B) Filtered orthologs: the number of RefSeq orthologous genes and introns derived from these data sets.

Intron class conversion

Orthologous mapping revealed that in all cases, orthologous mouse–rat and human–rodent introns—from the RefSeq data set—were either both U12 or both U2. A few cases were initially classified as instances of intron conversion. After close inspection, however, we realized that all of these

Table 2. Intron class and subclass frequencies in mammals

		Human	Mouse	Rat
U2	GT-AG	48,212 (98.9%)	44,817 (98.8%)	13,707 (98.7%)
	GC-AG	355 (0.7%)	330 (0.7%)	96 (0.7%)
	Other	184 (0.4%)	218 (0.5%)	80 (0.6%)
	Total	48,751	45,365	13,883
U12	GT-AG	131 (69.7%)	128 (71.9%)	36 (78.3%)
	AT-AC	51 (27.1%)	47 (26.4%)	9 (19.6%)
	Other	6 (3.2%)	3 (1.7%)	1 (2.2%)
	Total	188	178	46

cases could be explained either by misprediction of the intron boundaries or by splice sequence patterns slightly off consensus. (See Supplemental materials for the cross-species alignments at the intron boundaries of all predicted U12 introns). Remarkably, therefore, not one single convincing case of U12 to U2 conversion or vice-versa has occurred since the divergence of the human and rodent lineages. To investigate whether conservation of intron class extends beyond the mammalian lineage, we have mapped the 412 human, mouse, and rat U12 introns from Table 2, which correspond to 202 unique orthologs, into the chicken genome. The mapping was obtained by comparing, using exonerate (G. Slater, unpubl.), the two exons harboring the intron against the chicken genome sequence (see Methods). A total of 38 mammalian U12 introns were unequivocally mapped into the chicken genome. (See Supplemental material for cross-species alignments at the intron boundaries of the mammalian U12 introns mapped into the chicken genome). The 38 chicken introns had the typical donor-site sequence of U12 introns, and 36 had the typical U12 branch point. In the other two cases, sequences reminiscent of the U12 branch point could still be found, although departing clearly from the consensus. Since these two cases are both of the GT-AG U12 subtype, it is tempting to speculate that they may correspond to intermediates in the interconversion pathway between U12 and U2 introns. Against this hypothesis, however, is the fact that no strong polypyrimidine tract, suggestive of U2 function, can be found upstream of the acceptor site. With the exception of these two cases, the branch-point sequence was extremely conserved between mammals and chicken, showing no more than two mismatches, but often being identical. The position of the branch point has also been conserved; with only one exception, the larger displacement observed was of 4 nucleotides. These results strongly argue that U2 and U12 introns have evolved independently, at least since the split of mammals and birds.

Subtype switching

Although subtype switching between GT-AG and AT-AC U12 introns has been documented (Burge et al. 1998), we have not found any such case within rodents, between human and rodents, or between mammals and chicken in our set of U12 orthologous introns. It appears that this phenomenon occurs at a very slow rate over evolutionary time (see cross-species alignments of orthologous U12 introns in the Supplemental material).

Within U2 introns, on the contrary, switching between GC-AG and GT-AG subclasses, and vice-versa, is not unusual. Table 3A lists the pairwise frequency of subtype switching within U2 introns, and subtype distribution within orthologous mammalian triads. Because of the limited number of cases available in the RefSeq collection, we have ignored chicken genes in this analysis. A total of 190 of the 290 human (66%) and 289 mouse (66%) GC-AG introns are conserved in both species. Similar proportions are observed between human and rat. Within rodents, 60 of the 68 mouse (88%) and 67 rat (90%) GC-AG introns are conserved in both species. The availability of orthologous introns from three organisms allows the investigation of the dynamics of subtype switching within U2 introns (see Table 3B). We have divided GC-AG introns' orthologous triads into (1) "ancient"; the intron is GC-AG subtype in the three species, and thus it is likely to predate the split of human and rodents; (2) "modern"; the intron is GC-AG subtype in either human or rodents. Because of the lack of a reference out-group, however, we cannot distinguish here those ancient GC-AG introns that have reverted to GT-AG in one of the two lineages from those modern GC-AG introns that have arisen in one of the lineages; and (3) "recent"; the intron is of GC-AG subtype only in one of the rodent species. The most parsimonious hypothesis is that the switch to GC-AG has occurred after the split of mice and rats.

According to this classification, 47% (45) of the GC-AG introns are ancient, 36% (34) are modern, and 14% (13) are recent. Because human introns act as a reference out-group, we can establish (under the most parsimonious hypothesis) the direction of the GT/GC switch between mouse and rat orthologous introns. Although the numbers are too small to draw definitive conclusions, we observe more GT to GC than GC to GT substitutions (13 vs. 3). This is obviously mostly due to the overwhelmingly larger number of GT-AG than GC-AG introns, but indicates that switching from GT to GC in the donor site of U2 introns is not completely unfavorable. In this regard, it is interesting to note that GC-AG introns' exhibit a stronger and less variable do-

Table 3. Observed cases of U2 subtype switching within mammals

(A) Orthologous pairs

	GT, GT	GC, GC	GC, GT	GT, GC
human/mouse	38,922	190	100	99
human/rat	11,693	61	33	23
mouse/rat	8441	60	8	7

(B) Orthologous triads

Human	Mouse	Rat	Occurrences
	"ancient" GT-AG		
GT	GT	GT	7784
	"ancient" GC-AG		
GC	GC	GC	45
	"moderate" GC-AG		
GC	GT	GT	23
GT	GC	GC	11
	"recent" GC-AG		
GT	GT	GC	8
GT	GC	GT	5
	"ancient" GC-AG, "recent" GC → GT		
GC	GC	GT	2
GC	GT	GC	1
		Total	95

(A) Orthologous pairs: occurrence of donor site dinucleotide pairs at intron boundaries of orthologous intron pairs. For instance, we have found 65 instances in which the orthologous donor site is GC in human and GT in mouse.
(B) Orthologous triads: occurrence of donor site dinucleotides at intron boundaries in orthologous intron triads. For instance, we have found 23 cases in which the donor site is GC in human, but GT in both mouse and rat.

nor-site sequence than GT-AG introns (Fig. 1). Indeed, the information content of GC-AG donor sites is 12.4, while that of GT-AG donor sites is only 8.2. Probably, the substitution GT→GC, less favorable energetically, needs to be compensated by stronger complementarity in the rest of the site. Indeed, while GC-AG introns make up only 0.7% of all U2 introns (see Table 2), when considering only those U2 introns whose donor-site sequence is the perfect complement to the U1 snRNA 5′ end sequence ([AGC]AG/G[CT]AAGT), then, the percentage of GC-AG introns rises to 11.35% (317 of 2792).

Comparison of splice site sequence patterns

We have investigated here whether the splice-site sequence patterns have changed appreciably since the mammalian and avian split. One way to investigate the variation is to visually compare pictograms or logos (Fig. 1) obtained from collections of sites from different species, derived from the Ensembl database. To facilitate this task, we have extended sequence pictograms into comparative pictograms. In these, the nucleotide distributions of the two species at each position are represented side by side, and the ratio of the nucleotide proportions indexes a range of colors from green to red, indicating nucleotide overrepresentation in one of the two species (see Methods and Supplemental material). Figure 2 shows the comparative pictograms for mouse and rat, human and mouse, and human and chicken. For reference, we have also computed them for human and zebrafish and human and fly. As it is possible to see, comparative pictograms suggest that splice sequence patterns are largely homogeneous within tetrapoda (the pictograms are mostly yellowish), but noticeably distinct from those of other vertebrate and invertebrate taxa. Statistical analysis in which we have explicitly computed the distances between splice-site sequence patterns, using a variety of methods, supports this interpretation (see Supplemental material).

Sequence conservation of orthologous U2 splice sites

In this section, we investigate sequence conservation at orthologous splice sites. Here, we have used the HMRG set of curated mammalian–avian orthologous introns (Methods). In two ways, Figure 3 displays comparisons of orthologous splice sites, the percentage of sequence identity at each nucleotide position in the splice sites and at an intronic region 10 nucleotides long adjacent to the sites. Identity has been computed after aligning the orthologous splice-site sequences at the intron boundaries. Because these alignments are ungapped, the characteristic geometric decay of conservation within the intron observed for mouse–rat and for human–rodent comparisons is suggestive of significant sequence conservation between orthologous introns at this phylogenetic distance. In contrast, for mammalian and chicken comparisons, the ungapped alignment shows an almost abrupt decay right after the splice site—very similar to that observed when comparing unrelated sites.

To investigate what fraction of sequence conservation in splice sites is due to splicing function, we computed background sequence conservation between pairs of (randomly chosen) non-orthologous sites. As expected, background identity is ~25% outside of the splice signals. Within the splice signals, background conservation at each position roughly correlates with the information content at that position. Interestingly, at the acceptor site, it exhibits a bimodal shape—consistent with the polypyrimidine tract appearing at two different preferential locations. There is also a slow decay of background conservation upstream of the acceptor site—suggesting that the boundaries of this site are not precisely defined.

As shown in Figure 3, orthologous splice-site sequences are more conserved than expected solely from their role in splicing. Interestingly, this additional conservation is larger than that obtained at adjacent intronic sites for mammalian–chicken comparisons, but not for human–rodent and mouse–rat comparisons (Fig. 3, bottom). The abrupt decay of background conservation right after the donor site allows us to quantify this observation at these sites. This is less obvious in acceptor sites, because their boundaries are not as sharply defined. Indeed, we have computed the average sequence identity in the four rightmost intronic positions of the donor site (positions +3 to +6 in Fig. 1), and at four adjacent positions outside of the site (+7 to +10). The values of background conservation in these two regions are ~50% and 26%–27%, respectively, for all pairs of species. For mouse–rat orthologous comparisons, the values are 89% and 76%, respectively, for human–mouse, 78% and 53%, respectively, and for human–chicken, 62% and 31%, respectively. That is, conservation due to nonsaturation is smaller at the donor site than at adjacent positions (89 − 50 = 39% vs. 74 − 26 = 48%) for comparisons within rodents, similar for human–rodent comparisons (27% vs. 26%) and larger for human–chicken comparisons (12% vs. 4%). While it cannot be ruled out that this additional conservation reflects the existence of a small class of donor sites conserved beyond the generic consensus, a simpler explanation is that the reaching of saturation (understood here as the level of conservation at which orthologous sites are as conserved as unrelated sites, 27% identity at intronic sites, 50% at donor sites) is slower at sites under functional constraints. In the case of splicing, nucleotide substitutions at the splice sites may impair splice function. Thus, while the substitution process since the divergence of the mammalian and avian lineages has lead to almost complete saturation in proximal intronic sites (31% identity), donor sites (62% identity) are still far from saturation.

Discussion

Thanks to the availability of genome sequences for a number of mammalian and one avian species, we have been able to inves tigate the dynamics of the evolution of splice-site sequences in recent evolutionary times. Our results confirm that the splicing code is under evolution, albeit very slow. Indeed, while differences between overall splice-site sequence patterns correlate well with phylogenetic distance, they have remained largely homogeneous within tetrapoda, showing noticeable differences only at larger phylogenetic distances—such as those separating tetrapoda from fish.

Even though the splicing code appears to have remained quite constant within tetrapoda, our results also indicate that specific splice-site sequences may suffer significant changes during evolution and remain functional. Figure 3 displays the percentage of sequence identity at each nucleotide position across orthologous splice sites within rodents, between human and rodents, and within mammals and chicken. At all distances, orthologous splice-site sequences are more conserved than unrelated splice sites, but they have significantly diverged, showing an intermediate level of conservation between that of exon and intron sequences. The existence of additional sequences enhancing or repressing the recognition of the splice sites (for instance, see Caceres and Kornblihtt 2002; Cartegni et al. 2002; Black

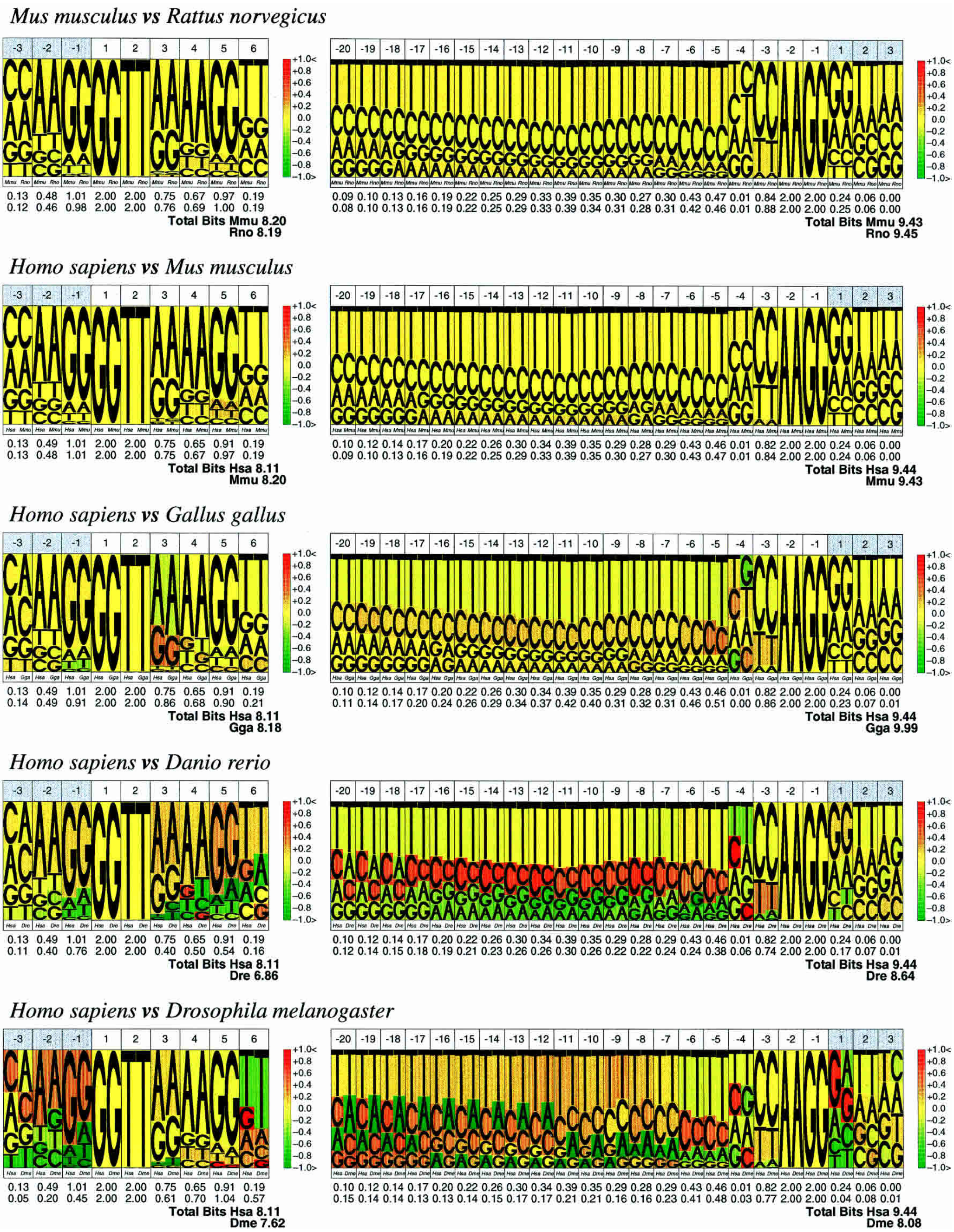


Figure 2. Comparative pictograms for donor and acceptor splice sites. Comparative pictograms of donor and acceptor sites for pairwise comparisons between species at different phylogenetic distances. At each position, the nucleotide distribution of the two species is displayed, the height of the letters corresponding to their relative frequency at the position. The color in the background of the letters indicates the underrepresentation (green) or overrepresentation (red) of a given nucleotide in the second species (*right*) with respect to the first (*left*).

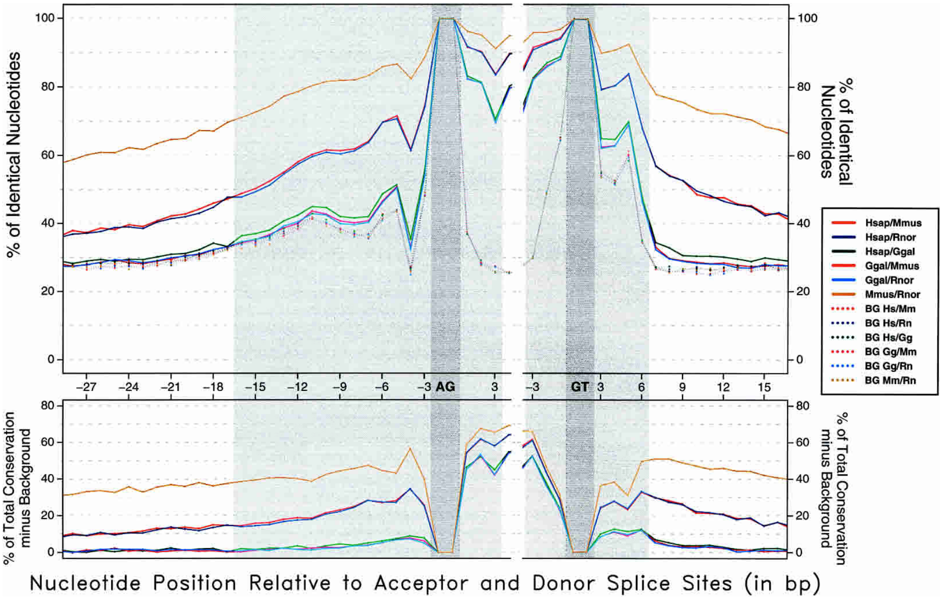


Figure 3. Sequence conservation level of orthologous GT-AG splice sites. Shaded gray areas correspond to the typical sequence span of splice-site signals. The average identity between the orthologous sequences is plotted across the splice signals (see Discussion). Background identity has been estimated from pairs of nonorthologous sites. (*Bottom*) The result of subtracting background conservation from total conservation.

2003) may partially explain the robustness of the exonic structure in front of changes in the splice-site sequences.

The greater conservation observed in mammalian chicken orthologous splice sites than in unrelated sites indicates that nucleotide substitution since the mammalian avian split has not yet reached saturation at these sites (estimated at ~50% identity at donor sites). At this phylogenetic distance, however, saturation has been reached at intronic sites, showing a level of conservation similar to that of unrelated sequences. This is the most likely explanation for the excess conservation over background observed in splice sites for comparisons between mammals and chicken, but absent in comparisons within mammals—where saturation has not been reached either at intronic sites.

In any case, the characteristic conservation of orthologous splice sites suggests that comparative prediction of splicing—through the modeling of the conservation in orthologous sites—could improve over methods based on the analysis of a single genome. Comparative prediction of splice sites could be particularly relevant to the prediction of alternative splicing—a problem still poorly solved—since it appears that a large fraction of alternative splicing events are conserved between related species, such as human and mouse (Thanaraj et al. 2003).

The availability of a large collection of orthologous intron sequences has also allowed us to investigate the evolutionary relationship between the minor U12 splicing class, and the major U2 class. Our results seem to indicate that U12 and U2 introns have evolved independently after the split of mammals and birds, since we have not been able to document a single convincing case of conversion between these two types of introns in our data sets. Certainly, because we have used a rather stringent criteria of U12 membership, it cannot be completely ruled out that such cases exist—maybe associated with dramatic changes in exonic structure, which our analysis cannot detect. On the other hand, although subtype switching between GT-AG and AT-AC U12 introns has been documented (Burge et al. 1998), we have not found any such case in our sets of U12 orthologous introns. In contrast, switching between the minor GC-AG and the major GT-AG subtypes within U2 introns is not unusual, and appears to be relatively well tolerated in introns with very strong donor sites. Comparison of orthologous introns has also allowed us to refine the sequences involved in the specification of the U12 introns (see Methods and Fig. 1). These sequences, while more conserved than signals involved in U2 intron specification, are more degenerate that previously thought.

Splicing remains an intriguing phenomenon. The results presented here, however, indicate that the increasing availability of sequences from genomes at different evolutionary distances will greatly contribute to the understanding of splicing, in particular, to understanding its history and its fundamental coding characteristics.

Methods

All of the statistical analyses were performed with the R package (Ihaka and Gentleman 1996; http://www.r-project.org/) using ad hoc scripts for the preparation of exploratory data analysis plots.

RefSeq genes and introns

Assembled chromosomal sequences and their associated annotations were downloaded from the UCSC Genome Browser (Kent et al. 2002; Karolchik et al. 2003; http://genome.cse.ucsc.edu/). The results described in this work were obtained on the assemblies listed in Table 1.

RefSeq genes interrupted with stop codons, or for which the amino acid sequence derived from the genomic coordinates had a difference of more than three amino acids in length or more than five gaps in the alignment when compared with the original amino acid sequence, were discarded. After this filtering step, 16,803 genes from the 21,744 annotated genes of the human HGv16 data set, 9734 genes from the 17,988 of the mouse MGSCv4, and 2783 genes from the 4877 of the rat RGSCv3.1 were retained.

Orthologous mammalian RefSeq introns

Gene sets

The set of homologous gene pairs was downloaded from the NCBI's HomoloGene database (Zhang et al. 2000; http://www.ncbi.nlm.nih.gov/HomoloGene/). From 369,338 homolog pairs, there were 46,522 pairs corresponding to human–mouse, human–rat, or mouse–rat orthologous genes. Redundancy was removed in order to keep only unique putative ortholog pairs. Only those gene pairs in which the two members were in the final gene set resulting after the filtering process above were taken into account. Ternaries of human, mouse, and rat genes were built when possible. Otherwise, the gene pairs were considered.

This process yielded 1283 human–mouse–rat triads. In addition, 4267 human–mouse ortholog pairs, 454 human–rat pairs, and 133 mouse–rat pairs were obtained. These numbers correspond to 6043, 5680, and 1847 unique RefSeq genes for human, mouse, and rat, respectively. When performing pairwise comparisons, the corresponding genes in the triads were included in the set of pairs. Thus, the resulting extended pair-wise sets contained 5550 human–mouse, 1737 human–rat, and 1416 mouse–rat pairs. All data sets, as well as graphical displays of sequence comparisons of the orthologous sequences are available from http://genome.imim.es/datasets/hmrg2004/.

Introns sets

We devised a protocol to extract orthologous intron pairs and triads from the above set of orthologous genes. First, all of the pairs of consecutive exons for each gene were aligned with t_coffee (Notredame et al. 2000; http://igs-server.cnrs-mrs.fr/cnotred/Projectshomepage/tcoffeehomepage.html) using default parameters against all of the exonic pairs from the corresponding orthologous genes. This step ensured that we were working with the most accurate set of orthologous introns, despite changes in the exonic structure of orthologous genes (such as missing exons due to misannotations or gaps in the assemblies). Second, the exonic structure of the gene was projected onto the alignments. Third, from orthologous gene pairs or ternaries, only those exon pairs in which all intron positions occurred at conserved positions in the alignment and the intron phases were conserved and retained. Plots on which the exonic structures have been projected onto the alignments can be accessed at http://genome.imim.es/datasets/hmrg2004/.

Orthologous HMRG introns

A set of human, mouse, rat, and chicken 1:1:1:1 confident orthologous introns was taken from International Chicken Genome Sequence Consortium (2004) (P. Bork and I. Letunic, pers. comm.). The set consisted of 1041 orthologous genes, totalizing 9110 orthologous introns. After mapping those genes into the annotations for the newer assemblies used in this analysis, 863 genes and 6524 introns remained in the four species orthologous set. The sequences 75 bp upstream and downstream of the signal core nucleotides (GT and AG for instance) were used in the orthologous splice-sites' sequence conservation analysis.

Intron class

U12 introns were searched, relying on the conserved donor-site sequence and the acceptor-site branch point. Mammalian introns were initially considered to be U12 if (1) they matched the motif [AG]TATCCTT (where [AG] means A or G) from position +1 at the donor splice site; and (2) they matched the motif TCCT T[AG]A[CT] at the region from -5 to -20 upstream the acceptor splice site. When looking for the U12 branch point, up to two mismatches were allowed, and the hit was accepted if at least one adenine was found in position 6 or 7 of the motif—to avoid branch point hits without biological sense. Visual inspection of introns orthologous to U12 introns, but which initially failed to meet this criteria, suggested that this initial definition is too stringent. Therefore, we searched only for the presence of a strong branch point signal at the appropriate location in orthologous introns. After inspection of all of those cases in which the two orthologous introns contain such a signal, we found a few additional cases in which the donor-site sequences strongly resemble the characteristic U12 donor site sequence, but failed to match the consensus above. Indeed, we have found that only the nucleotides at positions +2 (T), +3 (A), +4 (T), and +5 (C) within the intron are absolutely conserved in U12 donor-site sequences (TATC). Position +6, thought to be an invariable C (Burge et al. 1999), may also be a T, and positions +7 and +8 can actually be occupied by any nucleotide. This more degenerate pattern was the one used to identify chicken U12 introns, where, at most, a gap (in addition to one mismatch) was also allowed to match the branch-point consensus. These results, which help to characterize the sequences that define U12 introns, illustrate the power of comparative genomics to refine our knowledge of the functional sequences encoded in eukaryotic genomes.

Mapping of mammalian U12 introns into the chicken genome

DNA sequences of the exon-pairs delimiting each U12 intron were mapped into chicken genomic sequences using exonerate (http://www.ebi.ac.uk/guy/exonerate/). Only those alignments that preserved the mammalian splice site were taken into account. Introns obtained in that way were classified into U2/U12 classes following the same criteria as in the above section.

Comparison of splice site sequence patterns

We have quantified the different use of nucleotides in splice sites by different species and represent it by comparative pictograms. A comparative pictogram is a graphical representation of the nucleotide proportions observed in two different sets of aligned sequences. In this article, these sets are splice sites of different species and the proportions are calculated for every position along the splice site. As in sequence pictograms, the sizes of nucleotides scale with their observed proportions, but here the nucleotides of the two sets are put side by side to ease their comparison. Moreover, the background occupied by each nucleotide is colored with the ratio of the proportions (the relative risk). Further details are given in the Supplemental material.

We have further analyzed the different nucleotide usage in splice sites of different species by two kinds of comparisons as follows: (1) by building confidence intervals for the relative risks and counting how many of them include a ratio value of 1 (i.e., no difference of nucleotide usage), and (2) by assessing the site species dependence, that is, the extent to what the occurrences of the observed splice sites depend, statistically speaking, on the species to which they belong to. Further details are given in the Supplemental material also.

Acknowledgments

We thank the International Chicken Genome Sequencing Consortium for providing the genomic sequences, as well as the Rat and Mouse Consortia from past collaborations. We are particularly grateful to Ivica Letunic and Peer Bork for providing the set of HMRG orthologous introns with which some of the analyses were performed. Juan Valcárcel, Genís Parra, Eduardo Eyras, Webb Miller, David Haussler, Robert Baertsch, Chris Ponting, Alberto Roverato, Kim Worley, and two anonymous referees are gratefully acknowledged for advice and helpful comments. We also thank Òscar González for keeping the database mirrors up to date. Special thanks to Jan-Jaap Wesselink and Charles Chapple for their suggestions when proofreading this document. J.F.A. is supported by a predoctoral fellowship from the "Fundació IMIM" (Spain). This research is supported by grant BIO2000-1358-C02-02 from "Plan Nacional de I+D" (Spain), and grant ASD from the European Commission.

References

Birney, E., Andrews, T.D., Bevan, P., Caccamo, M., Chen, Y., Clarke, L., Coates, G., Cuff, J., Curwen, V., Cutts, T., et al. 2004. An overview of Ensembl. *Genome Res.* **14:** 925–928.

Black, D.L. 2003. Mechanisms of alternative pre-messenger RNA splicing. *Annu. Rev. Biochem.* **72:** 291–336.

Burge, C.B. and Karlin, S. 1998. Finding the genes in genomic DNA. *Curr. Opin. Struct. Biol.* **8:** 346–354.

Burge, C.B., Padgett, R.A., and Sharp, P.A. 1998. Evolutionary fates and origins of U12-type introns. *Mol. Cell* **2:** 773–785.

Burge, C.B., Tuschl, T., and Sharp, P.S. 1999. Splicing precursors to mRNAs by the spliceosomes. In *The RNA world* (eds. R.F. Gesteland et al.), pp. 525–560. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.

Burset, M., Seledtsov, I., and Solovyev, V. 2001. SpliceDB: Database of canonical and noncanonical mammalian splice sites. *Nucleic Acids Res.* **29:** 255–259.

Caceres, J.F. and Kornblihtt, A.R. 2002. Alternative splicing: Multiple control mechanisms and involvement in human disease. *Trends Genet.* **18:** 186–193.

Cartegni, L., Chew, S.L., and Krainer, A.R. 2002. Listening to silence and understanding nonsense: Exonic mutations that affect splicing. *Nat. Rev. Genet.* **3:** 285–298.

de Souza, S.J. 2003. The emergence of a synthetic theory of intron evolution. *Genetica* **118:** 117–121.

Dermitzakis, E.T., Reymond, A., Lyle, R., Scamuffa, N., Ucla, C., Deutsch, S., Stevenson, B.J., Flegel, V., Bucher, P., Jongeneel, C.V., et al. 2002. Numerous potentially functional but non-genic conserved sequences on human chromosome 21. *Nature* **420:** 578–582.

Dietrich, R.C., Incorvaia, R., and Padgett, R.A. 1997. Terminal intron dinucleotide sequences do not distinguish between U2- and U12-dependent introns. *Mol. Cell* **1:** 151–160.

Dietrich, R.C., Peris, M.J., Seyboldt, A.S., and Padgett, R.A. 2001. Role of the 3′ splice site in U12-dependent intron splicing. *Mol. Cell. Biol.* **21:** 1942–1952.

Fedorova, L. and Fedorov, A. 2003. Introns in gene evolution. *Genetica* **118:** 123–131.

Hall, S.L. and Padgett, R.A. 1994. Conserved sequences in a class of rare eukaryotic nuclear introns with non-consensus splice sites. *J. Mol. Biol.* **239:** 357–365.

Hare, M.P. and Palumbi, S.R. 2003. High intron sequence conservation across three mammalian orders suggests functional constraints. *Mol. Biol. Evol.* **20:** 969–978.

Helfman, D.M. and Ricci, W.M. 1989. Branch point selection in alternative splicing of tropomyosin pre-mRNAs. *Nucleic Acids Res.* **17:** 5633–5650.

Ihaka, R. and Gentleman, R. 1996. R: A language for data analysis and graphics. *J. Computat. Graph. Stat.* **5:** 299–314.

International Chicken Genome Sequencing Consortium. 2004. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature* (in press).

Jackson, I.J. 1991. A reappraisal of non-consensus mRNA splice sites. *Nucleic Acids Res.* **19:** 3795–3798.

Karolchik, D., Baertsch, R., Diekhans, M., Furey, T.S., Hinrichs, A., Lu, Y.T., Roskin, K.M., Schwartz, M., Sugnet, C.W., Thomas, D.J., et al. 2003. The UCSC genome browser database. *Nucleic Acids Res.* **31:** 51–54.

Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., and Haussler, D. 2002. The human genome browser at UCSC. *Genome Res.* **12:** 996–1006.

Kryukov, G., Castellano, S., Novoselov, S., Lobanov, A., Zehtab, O., Guigó;, R., and Gladyshev, V. 2003. Characterization of mammalian selenoproteomes. *Science* **300:** 1439–1443.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. 2001. Initial sequencing and analysis of the human genome. *Nature* **409:** 860–921.

Levine, A. and Durbin, R. 2001. A computational scan for U12-dependent introns in the human genome sequence. *Nucleic Acids Res.* **29:** 4006–4013.

Mattick, J.S. 2001. Non-coding RNAs: The architects of eukaryotic complexity. *EMBO Rep.* **2:** 986–991.

Notredame, C., Higgins, D.G., and Heringa, J. 2000. T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.* **302:** 205–217.

Pruitt, K.D., Tatusova, T., and Maglott, D.R. 2003. NCBI reference sequence project: Update and current status. *Nucleic Acids Res.* **31:** 34–37.

Rat Genome Sequencing Project Consortium. 2004. Genome sequence of the brown Norway rat yields insights into mammalian evolution. *Nature* **428:** 493–521.

Roy, S.W., Fedorov, A., and Gilbert, W. 2003. Large-scale comparison of intron positions in mammalian genes shows intron loss but no gain. *Proc. Natl. Acad. Sci.* **100:** 7158–7162.

Schneider, T. and Stephens, R. 1990. Sequence logos: A new way to display consensus sequences. *Nucleic Acids Res.* **18:** 6097–6100.

Sharp, P. and Burge, C. 1997. Classification of introns: U2-Type and U12-Type. *Cell* **91:** 875–879.

Smith, C.W. and Nadal-Ginard, B. 1989. Mutually exclusive splicing of α-tropomyosin exons enforced by an unusual lariat branch point location: Implications for constitutive splicing. *Cell* **56:** 749–758.

Thanaraj, T., Clark, F., and Muilu, J. 2003. Conservation of human alternative splice events in mouse. *Nucleic Acids Res.* **31:** 2544–2552.

Waterston, R.H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J.F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., An, P., et al. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420:** 520–562.

Xu, Q., Modrek, B., and Lee, C. 2002. Genome-wide detection of tissue-specific alternative splicing in the human transcriptome. *Nucleic Acids Res.* **30:** 3754–3766.

Yeo, G., Holste, D., Kreiman, G., and Burge, C. 2004. Variation in alternative splicing across human tissues. *Genome Biol.* **5:** R74.

Zhang, Z., Schwartz, S., Wagner, L., and Miller, W. 2000. A greedy algorithm for aligning DNA sequences. *J. Comput. Biol.* **7:** 203–214.

Zhu, W. and Brendel, V. 2003. Identification, characterization and molecular phylogeny of U12-dependent introns in the *Arabidopsis thaliana* genome. *Nucleic Acids Res.* **31:** 4561–4572.

Web site references

http://genome.imim.es/datasets/hmrg2004/; further supplemental materials for this study.

http://genome.cse.ucsc.edu/; UCSC Genome Browser, from which the human, mouse, rat and chicken feature annotations and genome assemblies used in this study were downloaded.

http://www.ensembl.org/; Ensembl Genome Browser, from which a larger set of human, mouse, rat and chicken gene annotation sets were retrieved.

http://www.ncbi.nlm.nih.gov/HomoloGene/; NCBI's HomoloGene database, from where initial RefSeq orthologous pairs were obtained.

http://igs-server.cnrs-mrs.fr/cnotred/Projectshomepage/tcoffeehomepage.html; a multiple sequence alignment package.

http://www.ebi.ac.uk/guy/exonerate/; a generic tool for sequence comparison.

http://www.r-project.org/; the R project for statistical computing.

Received August 4, 2004; accepted in revised form November 11, 2004.

4.3.3 ICGSC, *Nature*, 432(7018):695–716, 2004

PubMed Accession:

http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&list_uids=15592404&dopt=Abstract

Journal Abstract:

http://www.nature.com/cgi-taf/DynaPage.taf?file=/nature/journal/v432/n7018/abs/nature03154_fs.html

Supplementary Materials:

See Section 4.3.2 and the following URL:
http://www.nature.com/nature/journal/v432/n7018/suppinfo/nature03154.html

NOTE: Because of copyright restrictions, we cannot offer the article, please follow links for fulltext.

Chapter 5

Visualization Tools

If a picture is not worth a 1000 words,
to hell with it !
—Ad Reinhardt (note this is from the original Chinese quote
that "a picture is worth 10,000 words")

In this chapter the focus will shift towards the annotation and visualization process, describing those tools that permit to integrate data from different sources, including gene-prediction results, to present them to biologists in a comprehensive and comprehendible manner. These programs are intended to provide an overall view of our knowledge of a genomic region in a user-friendly interface, either static or interactive.

Before reporting our contribution to this field, we will place it in context with respect to other software. Therefore, a review of visualization tools provides the best frame to present our developments later. In the case of `gff2ps`, we have also participated in the cartography of the human, the fruit-fly and the mosquito genomes, and a special mention is deserved in the corresponding section.

5.1 A Review of Visualization Tools for Genomic Data

This section is not an in depth review, but an attempt to enumerate a broad spectrum of such software —ranging from the fully automated genome pipelines to the simple command-line programs—, and to highlight their application to comparative genome analyses. Programs are classified into three types: a) the database browsers, b) the annotation workbenches, that can be also used as browsers; and c) specific tools to visualize results from different sequence analysis, pointing the attention on those developed on top of alignment algorithms. We will not deal here with the libraries of code that contain programs or functions to plot data in any of the aforementioned classes, because they are of interest mostly to advanced users and computer specialists —for instance, `bioTk` [Searls, 1995], `bioWidgets` [Fischer *et al.*, 1999], the `Bioperl` Toolkit [Stajich *et al.*, 2002] or the Generic Model Organism Project (`GMOD`, see page 214, on Web Glossary).

5.1.1 Database browsers

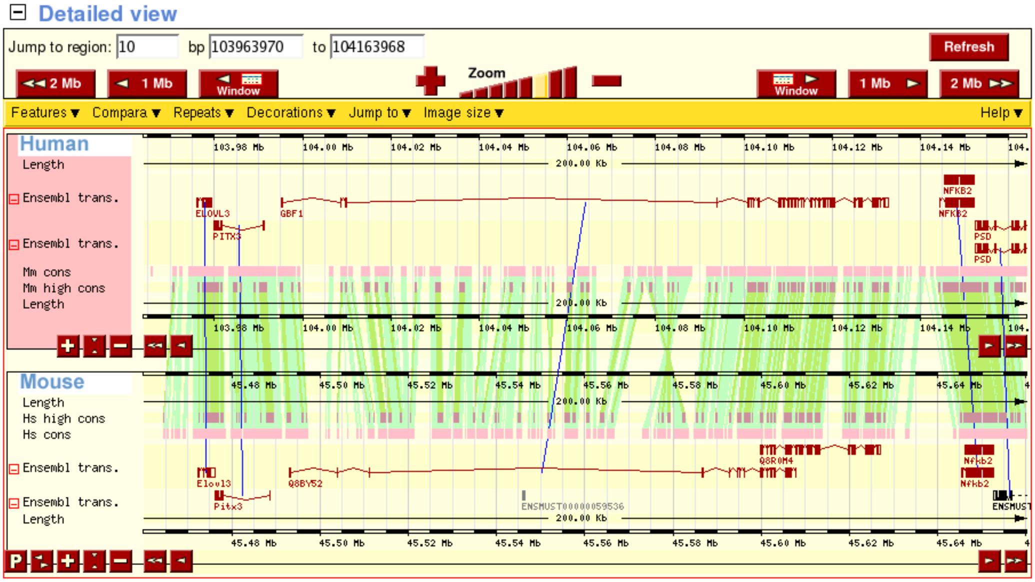
A first entry point to the visualization of genomic analyses can be any of the web front-ends developed to publish genome annotations. For example, the ones offered by databases of species- specific genome projects, such as the *Saccharomyces cerevisiae* SGD [Christie *et al.*, 2004], the *Caenorhabditis elegans* WORMBASE [Harris *et al.*, 2004], the *Drosophila melanogaster* FLYBASE [The `FlyBase` Consortium, 2003], the mouse MGD [Bult *et al.*, 2004], the *Arabidopsis thaliana* TAIR [Rhee *et al.*, 2003], and so on. The expected evolution of these of interfaces was to summarize all the information under a unified graphical schema as the number of species being sequenced increased —as done in the `euGenes` [Gilbert, 2002], the Generic Genome Browser (`Gbrowse`, Stein *et al.* 2002) and the `GeneDB` [Hertz-Fowler *et al.*, 2004] systems.

The best example of such evolution is ACEDB [see page 213, on Web Glossary; Durbin and Thierry-Mieg, 1993; Eeckman and Durbin, 1995], a seminal genome database system developed since 1989 and originally tailored for the *C. elegans* genome project. The tools in it have been generalized and are now used in a variety of organism-specific databases as diverse as bacteria and eukaryotes [Walsh *et al.*, 1998]. Specialized displays for managing and publishing genomic data are available through its well-set-up graphical user interface. Two remarkable implementations are the `AceBrowser` [Stein and Thierry-Mieg, 1998] and `Jade` [Stein *et al.*, 1998] programs.

There has already been a worldwide effort to centralize all the information about sequenced genomes. The best examples are the three fully established whole-genome browsers: the NCBI MAP VIEWER [see page 215, on Web Glossary; Wheeler *et al.*, 2005], the UCSC GENOME BROWSER [see page 216, on Web Glossary; Karolchik *et al.*, 2004] and the ENSEMBL system at the Sanger Institute and the EBI [see page 213, on Web Glossary; Birney *et al.*, 2004a]. All three browsers present by default a set of "in-house" and/or contributed gene- finding predictions from different programs. This is an on-going effort and predictions are recomputed for each newly released assembly. However, only the UCSC and ENSEMBL systems distribute predictions fully-based on the comparative genomics approaches. In what follows, we briefly review these three main genome gates.

The NCBI MAP VIEWER shows ab initio gene models generated by `Gnomon` [NCBI, 2003], a heuristic tool able to find the maximal self-consistent set of transcript and protein alignments to genomic data. Other programs like, for instance, `GenomeScan` [Yeh *et al.*, 2001], use this information to parameterize the constraints for an underlying HMM-based gene prediction model. The browser is focused to display genome assemblies using sets of synchronized chromosomal maps, but also features tables of genetic loci in homologous segments of DNA between human and mouse —the so called Human-Mouse Homology Maps—, and has links to HOMOLOGENE, a database of curated and calculated gene homologues.

The ENSEMBL system can display simultaneously different sets of annotated features and predictions from several gene-prediction tools embedded in the ENSEMBL annotation pipeline (see for instance, Figures 1.4 and 5.1). An interesting feature of the ENSEMBL system is the inclusion of external data through a Distributed Annotation System (`DAS`, Dowell *et al.* 2001) server, which, on user demand, dynamically links third-party annotations to the genomic sequence under study. The `SGP2` [Parra *et al.*, 2003], `Twinscan` [Korf *et al.*, 2001], and `SLAM` [Alexandersson *et al.*, 2003] gene annotation tracks, for instance, can be easily included in the current view by switching on the corresponding check box in the '*DAS*



Jump to region: 10 bp 103963970 to 104163968

Figure 5.1: **Human *GBF1* loci genomic region and its counterpart in mouse.** Detailed view of the human/mouse homology block at the *GBF1* loci (human chromosome 10, between 103963970 bp and 104163968 bp) as shown by the **MultiContig View** page on ENSEMBL. Orthologous genes are connected by a blue line. Pink boxes represent the homologous regions between both species projected into each sequence. Those homology hits are connected by green shaded regions. Differences at sequence level, such as insertions/deletions and inversions, are easily spotted with that green shading.

sources' drop-down menu. A syntenic regions navigation tool is available at ENSEMBL (see upper right panel from Figure 1.4 and Clamp *et al.* 2003). It was initially developed for human-mouse comparisons but it has been extended to include further species comparisons, i.e. rat, chicken, fruit-fly and so on. An example of the MultiContig viewer is shown in Figure 5.1.

Finally, gene-predictions can also be retrieved from the UCSC browser by switching on the appropriate options in the drop-down menus from the navigation form. In addition, the UCSC GENOME BROWSER features a novel database named ZOO, on which analyses made over a set of homologous targeted genomic sequences from 12 species [Thomas *et al.*, 2003] are published. Furthermore, depending on which genome is being browsed, the annotated gene features can be combined with the results of a mixture of whole-genome precomputed alignments from BLAT [Kent, 2002], BLASTZ [Schwartz *et al.*, 2003b], WABA [Kent and Zahler, 2000], and/or Exofish ecores [Jaillon *et al.*, 2003].

Current genome browsers, however, lack the ability to clearly represent information across genomes. A multiple species genome browser system should be able to represent many-to-many genomic alignments as an alignment among genomes. Moreover, it is difficult for most systems to develop a representation that natively compares whole-genomes and not only targeted regions. In this regard, the K-Browser [Chakrabarti and Pachter, 2004] has been designed around two principles: genome symmetry —every genome contains useful information, thus a browsing solution should not limit the ability to navigate within or across genomes—; and genome homology —related genomes have evolved from a common ancestor and these evolutionary relationships should be accurately reflected in both the representation and the visualization of information. The `K-Browser` takes as input a specific region in a specific genome and produces a set of images that succinctly represents the requested region and all orthologous regions. It can also provide the underlying multiple alignments.

5.1.2 Annotation workbenches

A myriad of sequence annotation workbenches have been developed during the last decade, but only a few have taken into account the comparative genomics perspective into their design. In this regard, it is worthwhile to cite `Alfresco` [Jareborg and Durbin, 2000], `genomeSCOUT` [Suter-Crazzolara and Kurapkat, 2000], `ERGO` [Overbeek *et al.*, 2003], `Theatre` [Edwards *et al.*, 2003], and `FamilyJewels` [Brown *et al.*, 2002]. Developed since the mid-nineties, these workbenches established the basis of modern annotation tools such as `Artemis` [Rutherford *et al.*, 2000] and `Apollo` [Lewis *et al.*, 2002]. The latter provides a human-mouse synteny panel that allows the user to compare and edit annotations for these two species. The `Artemis Comparison Tool` (`ACT`), based on the `Artemis` implementation, displays the results of a `BLASTN/TBLASTX` search along the sequence with the corresponding annotations. These tools are mainly employed by human curators for the re-annotation labour necessary to improve the raw annotations from automated pipelines. In this regard, the `Otter` annotation system [Searle *et al.*, 2004] extends the ENSEMBL database schema to integrate manual annotations by exchanging data in `XML` format between machines and allowing multiuser annotation. Two annotation tools have `Otter` client support, `Apollo` and `Otter/Lace`. `Otter/Lace` is a `perl` wrapper round the `AceDB` annotation editor, and it is currently used by the Human and Vertebrate Annotation (HAVANA) group curators at the Sanger Center. A review of several annotation browsers from the end-user viewpoint can be found in Fortna and Gardiner [2001].

5.1.3 Tools for visualizing alignments

Despite the trend to move from the pair-wise sequence comparison tools (two species) to the comparison of multiple sequences (many species) [Miller, 2001], there is still a niche for pair-wise comparison tools. The main reason is that such one-to-one alignments provide an informative comparison, but with the lowest complexity of interpretation.

Pair-wise comparisons can be done in several ways. A `dot-plot` or comparison matrix simultaneously displays all the structures in common between two sequences [Fitch, 1966; Gibbs and McIntyre, 1970]. In this, the conserved, repeated or inverted repeated segments are clearly visualized. Accordingly, `dot-plot` like diagrams have been extensively used to define the conserved segments of large genomic sequences, and also to explore the repeat-rich regions [Waterston *et al.*, 2002]. These conserved segments can be further analyzed with, for instance, the `PiP`-like tools described below. Among the pair-wise tools, one can cite `DIAGON` [Staden, 1982], `LFASTA` [Pearson and Lipman, 1988], `Lav` [Schwartz *et al.*, 1991], `Blixem` [Sonnhammer and Durbin, 1994], `Dotter` [Sonnhammer and Durbin, 1995], `Laj` [Wilson *et al.*, 2001], `GenoPix2D` [Cannon *et al.*, 2003], or `NOPTALIGN` [Smoot *et al.*,

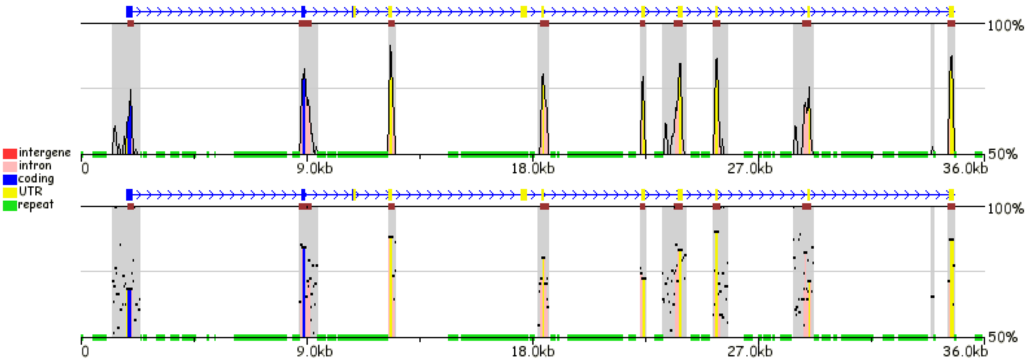


Figure 5.2: **A comparison of PiP-plots versus Smooth-plots.** Sequence between 95992 kb and 96028 kb from chromosome 8 was compared against its homologous mouse genomic sequence using the `zPicture` web server [Ovcharenko *et al.*, 2004a]. The same underlying alignment, computed with `BLASTZ` [Schwartz *et al.*, 2003b], is visualized as a pip-plot in the upper panel and as a smooth-plot in the bottom one, emulating the output from `PipMaker` [Schwartz *et al.*, 2000] and `VISTA` [Mayor *et al.*, 2000] respectively. Pip-plots display all the short ungapped alignments as black horizontal lines, while smooth-plots are constructed using, for each nucleotide, a 100 bp sliding window in which sequence identity is averaged. Boxes along the 100% identity baseline represent evolutionary conserved regions (ECRs), while those on the 50% baseline pinpoint the masked regions in which repetitive elements were found. *NM152416* gene structure (human hypothetical protein MGC40214) is depicted above the identity plots in both panels.

2004]. The `EMBOSS` suite [Olson, 2002; Rice *et al.*, 2000] provides several programs of this kind (`dottup`, `dotmatcher`, `dotpath` and `polydot`). The `gff2aplot` [Abril *et al.*, 2003] program falls within this software family. See Figure 5.8 on page 175 (Figure 1 on page 2478 of Abril *et al.* 2003), for examples of its output. Its major strength is to be independent of any alignment algorithm, as far as the input can be translated into the General Feature Format (`GFF`, see page 214, on Web Glossary). `TriCross` [Ray *et al.*, 2001], which extends the `dot-plot` concept to the simultaneous analysis of three sequences, renders the results in a three-dimensional Virtual Reality Modeling Language (VRML) representation.

Then again, those sequence comparisons can be represented in a more compact linear fashion. Several tools can be grouped here: `LAPS` (Local Alignment to PostScript, Schwartz *et al.* 1991), `LalnView` [Duret *et al.*, 1996], and `GenomePixelizer` [Kozik *et al.*, 2002]. The latter has been applied to visualize inter- and intra-chromosomal segmental duplications in genomic sequences [Cheung *et al.*, 2003; Estivill *et al.*, 2002].

Another class of programs, so called `PiP`-like because they produce Percentage Identity Plots, were designed to represent data from underlying sequence alignment algorithms. Basically, they consist in a compact display of the results of aligning one sequence to one or more sequences, where the positions (in the first sequence) and the score of the alignment segments are plotted, along with icons for features in the first sequence. `MUMmer` [Delcher *et al.*, 1999; Kurtz *et al.*, 2004], `PipMaker` [Schwartz *et al.*, 2000], `Multi-PipMaker` [Schwartz *et al.*, 2003a], `VISTA` [Mayor *et al.*, 2000], `CGAT` [Lund *et al.*, 2000], and `SynPlot` [Göttgens *et al.*, 2001], are among these tools. They do not fit into the gene-prediction paradigm *sensu strictu*; in any case, they have proven their potential in finding and/or refining protein-coding regions [Jang *et al.*, 1999; Pennacchio *et al.*, 2001; Reisman *et al.*, 2001; Tompa, 2001; Toyoda *et al.*, 2002; Wilson *et al.*, 2001], as well as the conserved non-coding sequences around them which may play a role in gene expression [Dubchak *et al.*, 2000; Gilligan *et al.*, 2002; Göttgens *et al.*, 2000, 2001; Hardison, 2000; Hardison *et al.*, 1997; Loots *et al.*, 2000; Oeltjen *et al.*, 1997; Ovcharenko and Loots, 2003b]. They even have been found useful in the analysis of the distribution of repetitive sequences [Chiaromonte *et al.*, 2001; Yuhki *et al.*, 2003]. See Figure 5.2 for an example of what can be done with these tools.

These programs have been reviewed in a number of occasions [Frazer *et al.*, 2003; Pennacchio and Rubin, 2001; Pennacchio, 2003; Pennacchio and Rubin, 2003; Thomas and Touchman, 2002; Ureta-Vidal *et al.*, 2003]. In Frazer *et al.* [2003], there is a good example of what can be achieved using those tools; it can be taken as a complete protocol describing how to retrieve the data sets, to prepare the sequences and complementary files, to compare them through the corresponding web browsers, and finally how to interpret their graphical outcomes. Two web servers have been deployed in an attempt to make those tools more interactive for the average user: the `ECR-Browser` (a navigation tool for Evolutionary Conserved Regions: Ovcharenko and Loots 2003a; Ovcharenko *et al.* 2004b) and `zPicture` (Ovcharenko *et al.* 2004a, and Figure 5.2). On the other hand, a comparison of the different alignment algorithm approaches behind some of those programs can be found in Ureta-Vidal *et al.* [2003]. `EnteriX` [Florea *et al.*, 2003] takes advantage of those principles to compare complete genomes of enteric bacteria. Nevertheless, the application of this algorithm to larger eukaryotic sequences, for instance to apply them in a whole-genome analysis, requires a large amount of computational resources. One drawback of these tools is that their input often needs to be defined within conserved genomic segments, for instance, regions of synteny between chromosomes, because sequence rearrangements can dramatically distort the corresponding alignments.

Some tools have been specifically devised for the analysis of regulatory regions, although they can use a similar approach that the one described above for programs such as `PipMaker` or `VISTA`. `ReguloGram` visualizes the density of co-occurring cis-element transcription factor binding sites measured within a 200 bp moving window through phylogenetically conserved regions. Within a high-scoring region, the relative arrangement of shared cis-elements within compositionally similar binding site clusters can be depicted then with `TraFacGram` [both, `ReguloGram` and `TraFacGram`, were described in Jegga *et al.* 2002]. `ConSite` [Lenhard *et al.*, 2003; Sandelin *et al.*, 2004] is a graphical web application that takes advantage of the phylogenetic footprinting to report putative transcription factor binding sites situated in conserved regions and located as pairs of sites in equivalent positions in alignments between two orthologous sequences.

Apart from raw sequence genomic comparisons, one might be interested in examining the gene distribution among two or more species. One of the first approaches to this was the `Oxford Grid` [Edwards, 1991]. Coordinates for successive chromosomes of two species were drawn along two axes as in a `dot-plot`, homologous loci were then depicted as dots. Pair-wise similarity scores have also been used to estimate closer neighbour relationships when analyzing many genomes as a whole. Those results have been commonly represented as pie charts or Venn diagrams [Blaxter *et al.*, 2002; Wood *et al.*, 2002], but this leads to an static view of the sequence relationships. A more dynamic view is the one offered by the `SimiTri` tool [Parkinson and Blaxter, 2003], in which the simultaneous display and analysis of the similarity relationships of the dataset of interest, in example the complete proteome of an organism, relative to three other databases can be achieved.

5.1.4 Tools for visualizing annotations

One of the first graphic programs devoted to determine the function of nucleic acid sequences was ANALYSEQ [Staden, 1984b], and its focus on finding coding-exons. In this context it is also worth mentioning, the RSVP package [Searls, 1993]—in which sequence analysis algorithms were encoded using the POSTSCRIPT language, and thus, could in principle be performed by the printer.

Although not necessarily comparative based, several gene-prediction tools display graphical output either through a web server or as a standalone software. This graphical output generally consists in colored shapes corresponding to coding exons or other functional elements along the genomic axis. This approach was notably pioneered in X-windows systems by GeneModeler [Fields and Soderlund, 1990] as an standalone platform, and by XGRAIL [Uberbacher and Mural, 1991] as a network-based client-server architecture. In all these cases, the visualization capabilities are strongly tied to a particular gene finding algorithm. More general and algorithm independent visualization tools have been also developed. This task has been facilitated by the general acceptance of GFF format, and its derivatives (see page 214 from Web Glossary), as a standard for genomic features annotations. gff2ps [Abril and Guigó, 2000], for instance, displays GFF files assuming that the file itself carries enough formatting information. Additional flexibility comes from the customization files defined by the user, and also because of the POSTSCRIPT output and the ability to handle multiple page formats. Examples of its output can be seen on Figure 5.4 on page 160 (Figure 1 on page 744 of Abril and Guigó 2000). Those people looking for an interactive and extensible visualization program, should take a look to the GUPPY system [Ueno *et al.*, 2003], implemented over the Lua scripting language [Ierusalimschy *et al.*, 1996]. Finally, it is worth to cite Sockeye [Montgomery *et al.*, 2004], a three-dimensional Java-based application that has been developed recently to compactly display comparative analyses.

Initial developments of circular maps were devoted to draw restriction maps over plasmid sequences, then were applied to represent bacterial circular chromosomes. However, linear maps are more appropriate for visualizing genomic features, and for comparative studies in particular —as Tufte [2001] claims, any distortion when plotting data that will lead to misinterpretation should be avoided. Among the tools developed to visualize genetic maps one can cite gRanch [Wada *et al.*, 1997], mapmerge [Nadkarni, 1998], mapplet [Jungfer and Rodriguez-Tome, 1998], FitMaps+ShowMap [Graziano and Arus, 2002], NCBI's MapViewer [Wheeler *et al.*, 2002], or cMap [Fang *et al.*, 2003]. Applications to produce circular or linear representations of genomic features were provided by several software packages; such as GCG [Devereux *et al.*, 1984], Staden [Staden *et al.*, 2000], SRS [Etzold and Argos, 1993], SEALS [Walker and Koonin, 1997], or EMBOSS [Olson, 2002; Rice *et al.*, 2000]. Further examples of this kind of tools are GenomePlot [Gibson and Smith, 2003], GenoMap [Sato and Ehira, 2003], and ZoomMap+MappetShow [Barillot *et al.*, 1999].

Finally, it is worth to mention a set of visualization tools that are useful in a more specific analysis context. For instance, graph-based display using exons as nodes produces more compact pictures of alternative splicing exonic structures. This approach has been implemented in SpliceNest [Coward *et al.*, 2002; Krause *et al.*, 2002], and SplicingGraphs [Heber *et al.*, 2002]. Software that analyses the repeats distribution and composition on genomic sequences often includes a graphical interface, in which repetitive regions are linked by using straight lines or arcs. In this category one can find MiroPEATS [Parsons,

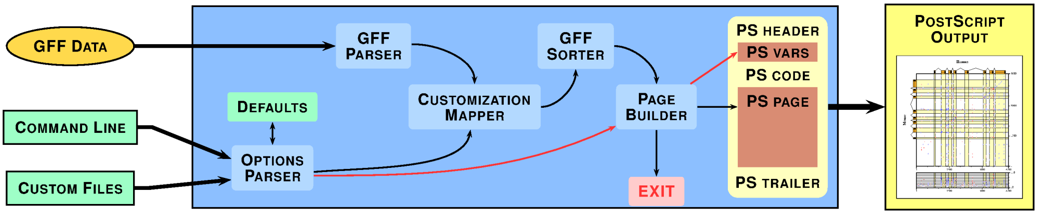


Figure 5.3: **Flow chart of internal main processes for `gff2ps` and `gff2aplot`.** Both tools were devised as standard `Unix` programs, they work as filters that process an input stream, in GFF, to produce an output stream, in POSTSCRIPT. Customization is provided by user-defined files or through command-line switches. Those settings are integrated with the input data to set a variables defining block and to bring forth the corresponding feature function calls in the page section of a POSTSCRIPT document. Such file is able to render the annotation plots thanks to specific POSTSCRIPT functions defined in its code section. The output document is self-contained, it has the data to plot and the commands to draw it.

1995], `REPUS` [Babenko *et al.*, 1999], `REPuter` [Kurtz and Schleiermacher, 1999], `Genome cryptographer` [Cleaver *et al.*, 2003], `Exact Match Annotator` [Healy *et al.*, 2003], `FORRepeats` [Lefebvre *et al.*, 2003], `GenomeComp` [Yang *et al.*, 2003], or `ADplot` [Taneda, 2004].

5.2 `gff2ps`: Visualizing Genomic Features

There are two major systems for representing graphic information on computers: raster and vector graphics. In raster graphics, an image is represented as a rectangular array of picture elements or pixels. Each pixel is represented either by its RGB color values or as an index into a list of colors. This series of pixels, also called a bitmap, is often stored in a compressed format. Since most display devices are also raster devices, displaying such a bitmap requires a viewer program to do little more than uncompress and transfer that bitmap to the screen. In a vector graphic system, an image is described as a series of geometric shapes. Rather than receiving a finished set of pixels, a vector viewing program, often also known as the interpreter, receives commands to draw shapes at specified sets of coordinates. In other words, it translates graphical objects into a virtual grid that is then projected in the corresponding raster device at a given fixed resolution. Although they are not as popular as raster graphics, vector graphics have one feature that makes them invaluable in many applications, they can be scaled without loss of image quality in the final rendering. This also means that once you generated an image you can zoom into any region of it to observe further details, which is done by the interpreter. To achieve the same with bitmaps requires to generate each zoom separately. This may not involve as much CPU time as needed by the vector graphics interpreter, but it is not efficient in storage space. Most of those arguments lead us to opt for a vector graphics programming language when developing most of our visualization tools, despite such systems do not have the same acceptance or support than a bitmap one. In any case, a vector graphic can be converted into a bitmap without loosing information while the other way around is not always true.

Introduced in 1985, PostScript is the name of a computer programming language developed originally by *Adobe Systems Incorporated* to communicate high-level graphic information to digital laser printers [Adobe S.I., 1999]. It is a flexible, compact, and powerful language both for expressing graphic images in a device-independent manner and for performing general programming tasks. The three most important aspects of the PostScript programming language are that it is interpreted, that it is stack-based, and that it uses a unique data structure called a dictionary. The dictionary mechanism gives the PostScript language a flexible, extensible base, and the fact that the language is interpreted and uses a stack model means that programs can be of arbitrary length and complexity. Since very little overhead is necessary to execute the programs, they can be interpreted directly from the input stream, which means that no memory restriction is placed on a PostScript program other than memory allocated by the program itself [Reid, 1996]. Those programming features make the PostScript language suitable for developing visualization tools in the genomic annotation field.

The combination of specific purpose PostScript-generating scripts previously implemented by me, along with the establishment of annotation interchange formats by the genome annotation community, such as GFF, led to the definition of the initial gff2ps draft. gff2ps was initially conceived in 1999 as a general drawing tool to represent gene-finding annotations from different sources. The program assumes that the GFF input itself carries enough formatting information. Genomic annotations have a hierarchical structure inherent to the biological features represented by them. For instance, a sequence may contain several genes, which are made of one or more exons, which are delimited by different signals, such splice sites and initiation or stop codons. Such structure is encoded in the GFF records by settling a fixed feature attribute on each field, i.e. the initial and terminal coordinates, a score, the group belonging to, and so on (see an example of the GFF record structure on page 214 from Web Glossary). gff2ps internal flow chart is depicted in Figure 5.3. Two main code blocks define this program: the gawk input filter and the PostScript drawing functions. The gawk code block is in charge of processing the GFF input records and the associated customization parameters, to produce specific PostScript-function calls for that data. Then, it embeds that piece of code in the PostScript document, which is by itself another code block.

Notable applications of gff2ps include the whole-genome annotation maps for several species —*Drosophila melanogaster* (Adams *et al.* 2000; see section 5.2.2 on page 161 and Figure 5.5), human (Venter *et al.* 2001; see section 5.2.3 on page 165 and Figure 5.6), the mouse chromosome 16 [Mural *et al.*, 2002], *Anopheles gambiae* (Holt *et al.* 2002; see section 5.2.4 on page 169 and Figure 5.7), and *Blochmannia floridanus* [Gil *et al.*, 2003]. Figure 3.8 on page 91 (Figure 2 on page 1142 of Guigó *et al.* [2003]) and bottom panel of Figure 5.10 are examples of using gff2ps in the comparative genomics context.

5.2.1 Abril and Guigó, *Bioinformatics*, 16(8):743–744, 2000

PubMed Accession:

http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&list_uids=11099262&dopt=Abstract

Journal Abstract:

http://bioinformatics.oupjournals.org/cgi/content/abstract/16/8/743

Program Home Page:

http://genome.imim.es/software/gfftools/GFF2PS.html

5.2.2 Adams *et al*, *Science*, 287(5461):2185–2195, 2000

PubMed Accession:

http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&list_uids=10731132&dopt=Abstract

Journal Abstract:

http://www.sciencemag.org/cgi/content/abstract/287/5461/2185

Companion Poster:

See Figure 5.5 and the following URLs:
http://www.sciencemag.org/feature/data/genomes/2000/drosophila.shl
http://genome.imim.es/references/genome_maps/2000_Science_v287_i5461_p2185_fig4_FlyGenome.ps.gz

NOTE: Because of copyright restrictions, we cannot offer the article, please follow links for fulltext.

5.2.3 Venter *et al*, *Science*, 291(5507):1304–1351, 2001

PubMed Accession:

http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&list_uids=11181995&dopt=Abstract

Journal Abstract:

http://www.sciencemag.org/cgi/content/abstract/291/5507/1304

Companion Poster:

See Figure 5.6 and the following URLs:
http://www.sciencemag.org/cgi/content/full/291/5507/1304/DC2
http://genome.imim.es/references/genome_maps/2001_Science_v291_i5507_p1304_fig1_HumanGenome.ps.gz

NOTE: Because of copyright restrictions, we cannot offer the article, please follow links for fulltext.

5.2.4 Holt *et al*, *Science*, 298(5591):129–149, 2002

PubMed Accession:

http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&list_uids=12364791&dopt=Abstract

Journal Abstract:

http://www.sciencemag.org/cgi/content/abstract/298/5591/129

Companion Poster:

See Figure 5.7 and the following URLs:
http://www.sciencemag.org/cgi/content/full/298/5591/129/DC2
http://genome.imim.es/references/genome_maps/2002_Science_v298_i5591_p129_fig1_MosquitoGenome.ps.gz

NOTE: Because of copyright restrictions, we cannot offer the article, please follow links for fulltext.

5.3 Software Developed for Comparative Analyses

5.3.1 gff2aplot: visualizing pairwise homology

gff2aplot was designed following the same principles as for gff2ps. Figure 5.3 illustrates the main internal processes flow chart for both tools. The problem to solve here was to integrate annotation information of two sequences being compared along with the pair-wise alignments obtained by other programs.

Due to the fact that each alignment software outputs alignments in their own format, it was decided to provide different filters to convert those alignment formats into a single interchange format. Such format was initially derived from GFF version 1, the so called aplot format. However, GFF version 2 provides enough flexibility to encode the alignment records into a more standardized way. Both alignment input formats, the aplot and the GFFv2, have been kept for backward compatibility in newer releases of gff2aplot. Use of an standardized input format permits to combine data from different alignment tools, or from different analyses made with the same tool—see for instance, right panel from Figure 5.8 on page 175 (Figure 1 on page 2478 of Abril *et al.* 2003)—, in order to compare them. An additional advantage of working with such filters to produce GFF-like records was the capability of visualizing that kind of data using gff2ps (as shown in Figure 5.10 lower panel).

Having that in mind, four programs have been implemented to date to complement gff2aplot, three perl scripts and another written in the C language. parseblast is a parser for the standard output from four of the BLAST program flavours available, say here NCBI-Blast [Altschul *et al.*, 1990, 1997], WU-Blast [Gish, 1996–2004], WebBlast [Ferlanti *et al.*, 1999] and MegaBlast [Zhang *et al.*, 2000]. blat2gff converts BLAT [Kent, 2002] output into GFF, while sim2gff does the same for SIM [Huang and Miller, 1991] output. The C program, ali2gff, processes SIM or Mummer [Delcher *et al.*, 1999] output to produce the GFF records for the alignment.

5.3.2 Abril *et al*, *Bioinformatics*, 19(18):2477–2479, 2003

PubMed Accession:

http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&list_uids=14668236&dopt=Abstract

Journal Abstract:

http://bioinformatics.oupjournals.org/cgi/content/abstract/19/18/2477

Program Home Page:

http://genome.imim.es/software/gfftools/GFF2APLOT.html

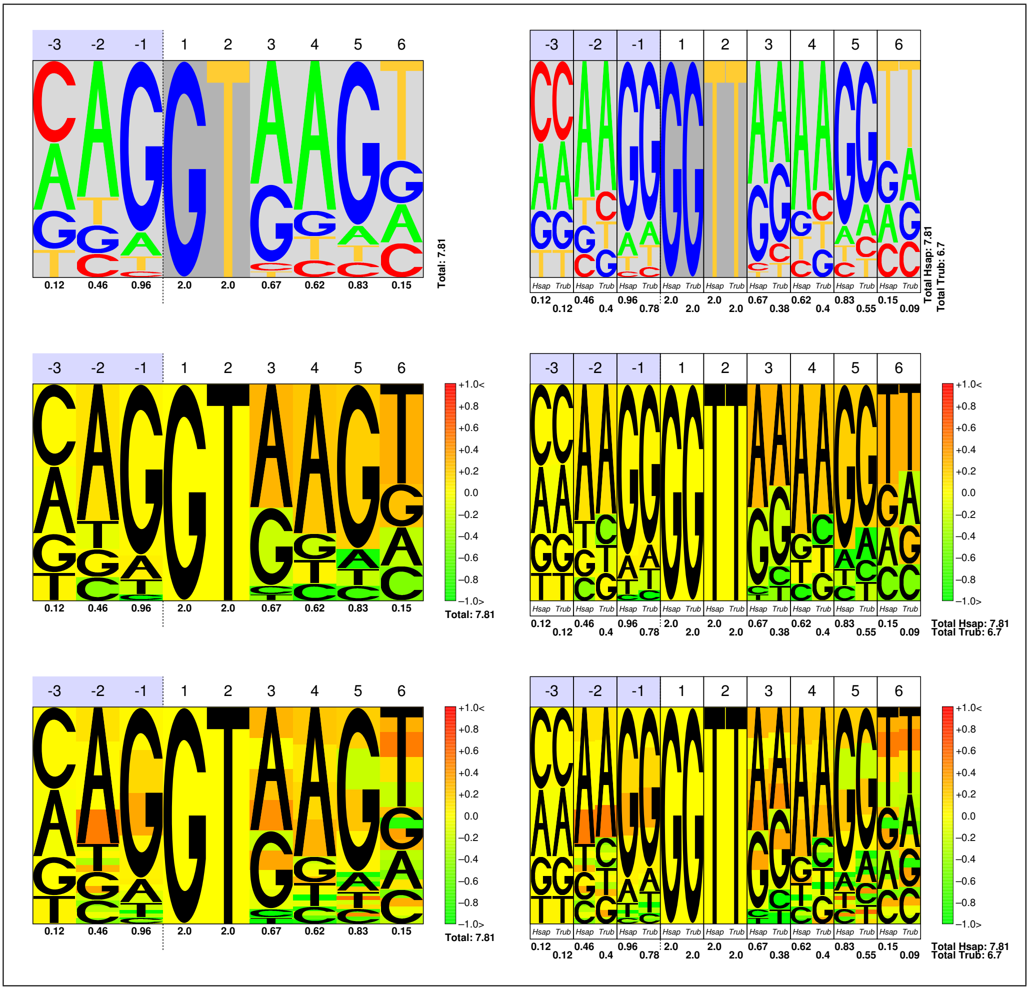


Figure 5.9: **Comparative pictograms.** We have initially developed `compi` to help in comparative analyses of splice sites. It can produce two kind of pictograms: the "standard" views, visualizing a pictogram for single species (left panels) and "comparative" views, currently set for pair-wise species matrices comparison (right panels). Depending on the input matrix, three different plots can be obtained, from top to bottom: the basic pictograms (with extra customizable layout), the Position-specific Scoring Matrices (PSMs) and the First-order Markov Models (FMMs) representations.

5.3.3 `compi`: Comparative pictograms

In sequence pictograms [Burge *et al.*, 1999]—which are analogous to sequence logos [Schneider and Stephens, 1990], the frequencies of the four nucleotides at each position along the signal—, the so called Position Weight Matrices [PWMs; Staden, 1984a, 1988; though the nowadays preferred term is Position-specific Scoring Matrices or PSMs] are represented by the heights of their corresponding letters. The information content (intuitively, the deviation from random composition) is computed at each position. It ranges from zero to two, with zero indicating random composition, and two indicating fixation of one nucleotide. The information content of the signal is the sum of the information content at each position. The larger the information content, the more conserved the signal (and, thus, more "informative": the smaller is the probability of finding it by chance). The relative entropy formula (also known as the Kullback-Leiber distance; Burge *et al*. 1999) is used to calculate the information content of the signal, as follows:

$$H_{signal} = \sum_{j=1}^{N} \sum_{i,j} P_{i,j} \log_2 \frac{P_{i,j}}{Q_i} \quad .$$

Where $N = length(signal)$, and $i \in \{A, C, G, T\}$. $P_{i,j}$ is the probability of finding nucleotide $i \in \{A, C, G, T\}$ in the j^{th} nucleotide of the signal, and Q_i is the probability of that nucleotide under the background distribution. By default, `compi` assumes the random distribution as background (so that, each $Q_i = \frac{1}{4}$), although other distributions can be provided by the user.

By inspecting the pictograms for two or more species, one tries to spot the different use, made by each of the species, of the nucleotides along the signal. This inspection, however, can become a difficult task for the following reasons. First, the differences in the size of each nucleotide can be difficult to observe as the two nucleotides are located in different pictures. Second, these differences are not quantified and thus we cannot assess with precision when a nucleotide is used more frequently in one of the species. Third, the assumption of marginal independence among the positions of the signal—implicit in PWMs—can hide relevant differences between species with regard to the dependencies between nearest neighbour positions along the splice signal.

We have tackled all three problems. First we have placed the nucleotides that occur in the same position, in the two species being compared, next to each other. Second we have calculated the ratio of the two relative frequencies (the odds) of each nucleotide in each position and represent the $\log_2$ of this ratio with a color code from green ($\log_2 \frac{1}{2} = -1$) to red ($\log_2 2 = 1$), where yellow is a ratio of 1 (0 in log-scale). The log-odds values of -1 and 1 work as saturation values and therefore, odds smaller than 0.5 or larger than 2 take green and red color, respectively. This color fills the rectangle defined by the nucleotide character and allows easy spotting of which nucleotides show a different occurrence between species. Third, we have extended the pictogram idea to represent first order dependencies between adjacent positions of the splice site—the so called First-order Markov Model (FMM). We have computed and represented the ratios of occurrence of each nucleotide with respect to the occurrence of every nucleotide in the previous positions. The representation has been implemented by splitting the rectangle defined by a nucleotide character in four equal rectangles, and filling out each of them with the color that corresponds to each of the ratios following a fixed order of A,C,G, and T. We shall refer to this representation as a *comparative pictogram* (`compi`). When rendering FMMs, the relative entropy at each position for each nucleotide is also weighted with respect to the occurrence of every nucleotide in the previous positions.

We split the task of producing the comparative pictograms in two, using separate `perl` scripts for each part. The first one computes nucleotide frequencies, ratios and First-order Markov dependencies from a set of sequences of fixed length. Then the matrices obtained

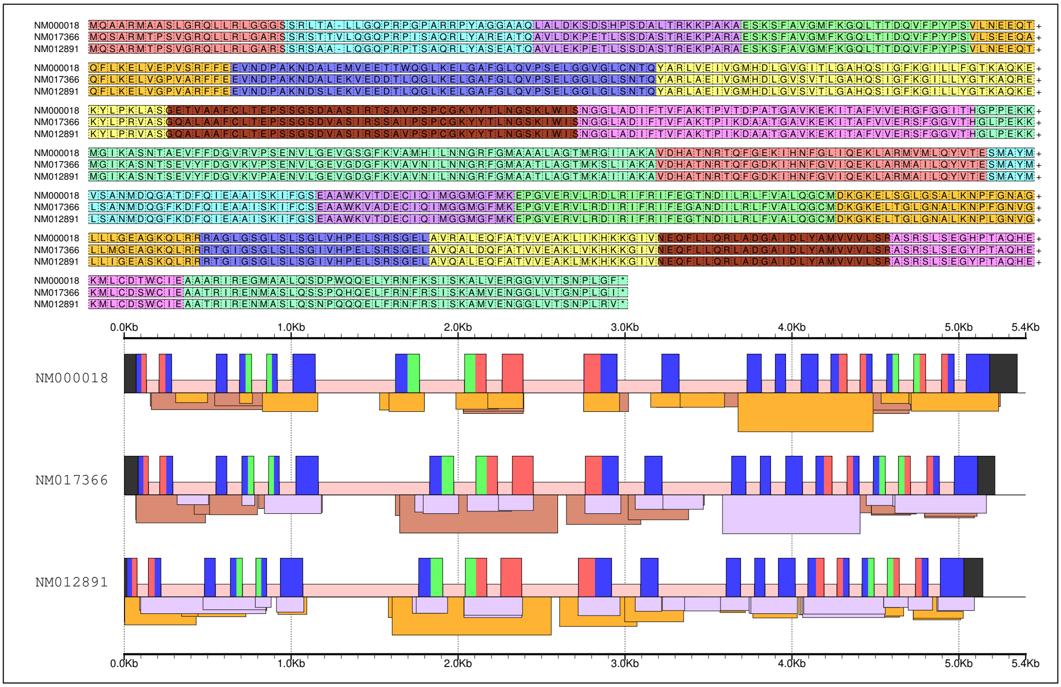

NM000018 MQAARMAASLGRQLLRLGGGSSRLTA-LLGQPRPGPARRPYAGGAAQLALDKSDSHPSDALTRKKPAKAESKSFAVGMFKGQLTTDQVFPYPSVLNEEQT+
NM017366 MQSARMTPSVGRQLLRLGARSSRSTTVLQGQPRPISAQRLYAREATQAVLDKPETLSSDASTREKPARAESKSFAVGMFKGQLTIDQVFPYPSVLSEEQA+
NM012891 MQSARMTPSVGRQLLRLGARSSRSAA-LQGQPRPTSAQRLYASEATQAVLEKPETLSSDASTREKPARAESKSFAVGMFKGQLTTDQVFPYPSVLNEEQT+
NM000018 QFLKELVEPVSRFFEEVNDPAKNDALEMVEETTWQGLKELGAFGLQVPSELGGVGLCNTQYARLVEIVGMHDLGVGITLGAHQSIGFKGILLFGTKAQKE+
NM017366 QFLKELVGPVARFFEEVNDPAKNDALEKVEDDTLQGLKELGAFGLQVPSELGGLGLSNTQYARLAEIVGMHDLGVSVTLGAHQSIGFKGILLYGTKAQRE+
NM012891 QFLKELVGPVARFFEEVNDPAKNDSLEKVEEDTLQGLKELGAFGLQVPSELGGLGLSNTQYARLAEIVGMHDLGVSVTLGAHQSIGFKGILLYGTKAQKE+
NM000018 KYLPKLASGETVAAFCLTEPSSGSDAASIRTSAVPSPCGKYYTLNGSKLWISNGGLADIFTVFAKTPVTDPATGAVKEKITAFVVERGFGGITHGPPEKK+
NM017366 KYLPRVASGQALAAFCLTEPSSGSDVASIRSSAIPSPCGKYYTLNGSKIWISNGGLADIFTVFAKTPIKDAATGAVKEKITAFVVERSFGGVTHGLPEKK+
NM012891 KYLPRVASGQALAAFCLTEPSSGSDVASIRSSAVPSPCGKYYTLNGSKIWISNGGLADIFTVFAKTPIKDAATGAVKEKITAFVVERSFGGVTHGLPEKK+
NM000018 MGIKASNTAEVFFDGVRVPSENVLGEVGSGFKVAMHILNNGRFGMAAALAGTMRGIIAKAVDHATNRTQFGEKIHNFGLIQEKLARMVMLQYVTESMAYM+
NM017366 MGIKASNTSEVYFDGVKVPSENVLGEVGDGFKVAVNILNNGRFGMAATLAGTMKSLIAKAVDHATNRTQFGDKIHNFGVIQEKLARMAILQYVTESMAYM+
NM012891 MGIKASNTSEVYFDGVKVPAENVLGEVGDGFKVAVNILNNGRFGMAATLAGTMKAIIAKAVDHATNRTQFGDKIHNFGVIQEKLARMAILQYVTESMAYM+
NM000018 VSANMDQGATDFQIEAAISKIFGSEAAWKVTDECIQIMGGMGFMKEPGVERVLRDLRIFRIFEGTNDILRLFVALQGCMDKGKELSGLGSALKNPFGNAG+
NM017366 LSANMDQGFKDFQIEAAISKIFCSEAAWKVADECIQIMGGMGFMKEPGVERVLRDIRIFRIFEGANDILRLFVALQGCMDKGKELTGLGNALKNPFGNVG+
NM012891 LSANMDQGFKDFQIEAAISKIFGSEAAWKVTDECIQIMGGMGFMKEPGVERVLRDIRIFRIFEGTNDILRLFVALQGCMDKGKELTGLGNALKNPLGNVG+
NM000018 LLLGEAGKQLRRRAGLGSGLSLSGLVHPELSRSGELAVRALEQFATVVEAKLIKHKKGIVNEQFLLQRLADGAIDLYAMVVVLSRASRSLSEGHPTAQHE+
NM017366 LLMGEAGKQLRRRTGIGSGLSLSGIVHPELSRSGELAVQALDQFATVVEAKLVKHKKGIVNEQFLLQRLADGAIDLYAMVVVLSRASRSLSEGYPTAQHE+
NM012891 LLIGEASKQLRRRTGIGSGLSLSGIVHPELSRSGELAVQALEQFATVVEAKLMKHKKGIVNEQFLLQRLADGAIDLYAMVVVLSRASRSLSEGYPTAQHE+
NM000018 KMLCDTWCIEAAARIREGMAALQSDPWQQELYRNFKSISKALVERGGVVTSNPLGF*
NM017366 KMLCDSWCIEAATRIRENMASLQSSPQHQELFRNFRSISKAMVENGGLVTGNPLGI*
NM012891 KMLCDSWCIEAATRIRENMASLQSNPQQQELFRNFRSISKAMVENGGLVTSNPLRV*
0.0Kb
1.0Kb
2.0Kb
3.0Kb
4.0Kb
5.0Kb
5.4Kb
NM000018
NM017366
NM012891
0.0Kb
1.0Kb
2.0Kb
3.0Kb
4.0Kb
5.0Kb
5.4Kb

Figure 5.10: **Merging exonic structure with coding sequence alignments.** Comparing the exonic structure of a set of orthologous genes (REFSEQ codes *NM000018*, *NM017366*, and *NM012891* in human, mouse, and rat respectively). At the protein level (top), splice sites were mapped over the amino acid alignment, and consecutive underlying exons were represented by alternating light and dark grey boxes. At the genomic level (bottom), the exonic structures are depicted along with the filtered best hits calculated from pair-wise `WU-TBLASTX` [Gish, 1996–2004] of comparisons of each sequence against the other two. The height of the boxes under the sequence axes correlates with the alignment score. The lower panel was obtained by `gff2ps` [Abril and Guigó, 2000].

are processed by a second script which generates POSTSCRIPT code specifically developed for the corresponding graphical representation of the matrices. This script can produce six different outputs, three "standard" (visualizing a pictogram for single species) and three "comparative" views (currently set for pair-wise species matrices comparison), which are shown in Figure 5.9. Computing the matrices outside the graphical program gives more flexibility to the user, who can preprocess matrices from other software to fit the input format of our tool (see page 213, on Web Glossary). This tool has been used to produce the pairwise pictograms shown in Figure 4.13 on page 134 (Figure 2 on page 116 of Abril *et al.* 2005).

5.3.4 Other developments

Several graphical procedures have been developed other than those shown until now, although many of them either are not finished enough to release to the community, or are quite specific for a given analysis to be really useful in another context. We are going to point out few of them in this section.

The need to combine the exonic structures along with sequence alignments at nucleotide or amino acid level, led to the development of the boxed alignments script for which an example is shown in Figure 5.10 upper panel. A more elaborated program developed in our group, named `exstral` (EXon STRucture over an ALignment, Castelo *et al.* 2004), produces a more quantitative output. However, its current text-based output lacks the integration achieved with the boxed alignments—for instance, to highlight subtle frame shifts in the exonic structure. The boxed alignments script generates a POSTSCRIPT plot. It will be interesting in the future to implement such kind of output into `exstral`.

As much important as writing procedures to analyze genomic data sets, is to choose an appropriate way to visualize the final results. The customization flexibility characteristic of `gff2ps` makes this tool useful to draw annotation features from different kind of analyses. Given a properly formatted input set of GFF records and taking the time to define an associated customization file or files, a researcher can obtain simple or complex representations of his annotations. It is then easy to apply those settings to a set of annotations for different sequences. Lower panel from Figure 5.10 shows an example of using `gff2ps` in a comparative genomics approach.

Chapter 6

Discussion

So easy it seamed once found, which yet unfound
most would have thought impossible

—John Milton

A central goal of genome analysis is the identification of all human genes. This task remains challenging, but is greatly aided by the near-complete sequence of the human genome [International Human Genome Sequencing Consortium, IHGSC, 2004], together with other improved resources (such as expanded cDNA collections, genome sequence from other organisms and better computational methods). The inventory of the best-defined functional components in the human genome—the protein coding sequences—is still incomplete for a number of reasons, including the fragmented nature of eukaryotic genes. The human gene number estimates, though, are coming closer to the real number of genes, as can be seen in Figure 6.1. To this end, there are several ongoing projects focusing on the definition of the precise catalog of human genes. One of those projects is the Vertebrate Genome Annotation (VEGA) database, a central repository for high quality, frequently updated, manual annotations of vertebrate finished genome sequences [Ashurst *et al.*, 2005]. The comparative sequencing program at the NIH Intramural Sequencing Center (NISC) aims to sequence and to analyze targeted genomic regions in multiple vertebrates [Thomas *et al.*, 2003]. The initial target of this project was a genomic segment of about 1.8 Mb on human chromosome 7q31.3 containing the gene encoding the cystic fibrosis transmembrane conductance regulator (*CFTR*) and nine other genes. Sequence clones for the orthologous genomic segments in multiple other vertebrates were obtained in order to perform an exhaustive comparative analysis of that region. The American National Human Genome Research Institute (NHGRI) launched a public research consortium, the ENCyclopedia Of DNA Elements (ENCODE) project [ENCODE Project Consortium, 2004], in September 2003, to carry out a project to identify all functional elements in the human genome sequence. The project is currently in its pilot phase, the evaluation of the procedures that can be applied cost-effectively and at high-throughput to accurately and comprehensively characterize large sequences. A set of 44 discrete regions—ranging in size from 0.5 to 2 Mb, that together constitute $\sim$1% of the human genome (30 Mb)—was chosen to represent a range of genomic features.

The unexpectedly low number of genes identified in the human genome raises again the

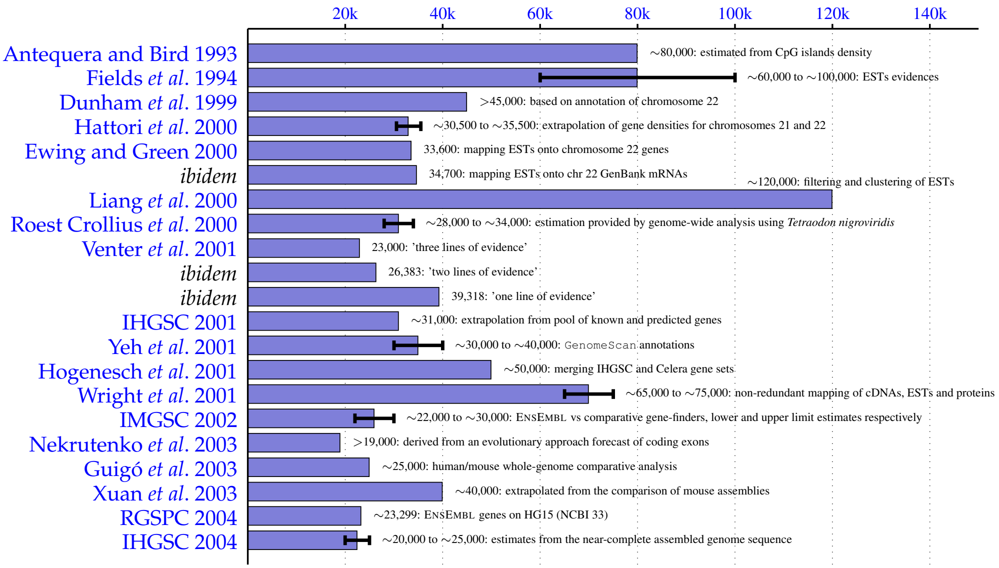


Figure 6.1: **Human gene number estimates in the genome era.** The figure depicts the number of human genes (blue bar) from various estimates, along with the references in which they were reported. It is worth to note that genes may produce more than one transcription unit or transcripts, which is not taken into account in this picture. Adapted from Harrison *et al.* [2002].

question of the source of an organism's complexity. One possible source is the greater structural complexity of the human genes, along with a higher level of regulation of those genes and the pathways in which they are involved. Another source are post-transcriptional modifications, more than 200 types are known and is predicted that three different modified proteins are produced for each human gene on average [Banks *et al.*, 2000]. Furthermore, alternative splicing of human genes might provide many more proteins per gene than in other organisms. Nevertheless, in Brett *et al.* [2002] they found similar levels of alternative splicing across species which argues against an overall increase in splicing as a source of increase in genome and organism complexity. Their data also suggested that a wide variety of gene products are further diversified by post-translational modifications. More recently, though, Pan *et al.* [2005] have provided evidence that at least 11% of human and mouse cassette alternative splicing events represent conserved exons that undergo species-specific alternative splicing. Such events have the potential to modulate frequently the structural and functional properties of proteins that are attributed to conserved domains. Therefore, they conclude that they could have an important role in the evolutionary differences between mammalian species. On the other hand, the recent identification of several types of ncRNAs, such as small nucleolar RNAs, microRNAs, guide RNAs and anti-sense RNAs, would significantly expand the complexity of the human genome [Storz, 2002]. Given the absence of a diagnostic open reading frame, a major question arises on how these genes can be identified. Novel evidences obtained by using high-density oligonucleotide arrays on different cell lines provide support for transcription outside well-characterized human exons [Kampa *et al.*, 2004]. Those transcribed regions, also known as transfrags, will provide a new view of the human transcriptome by mapping transcription to the genomic sequences.

One of the major obstacles towards the completion of the catalog of human genes is our inability to assess the reliability of the large number of computational gene predictions that have yet to be verified experimentally. Results described in Parra *et al.* [2003] demonstrate that through the comparison of related genomes, human and mouse in that example, and using the available comparative gene-finding tools, the false-positive rate can be reduced significantly, resulting in an improved catalog of vertebrate genes. Indeed, the experimental verification of a subset of those predictions provided evidence for at least 1000 previously non-confirmed genes [Guigó *et al.*, 2003]. The availability of another vertebrate species whose evolutionary position lies between mammals and fish would be of great utility to complete the vertebrates gene catalog. The success of these studies, suggests a new paradigm in high throughput genome annotation, in which gene predictions serve as the hypothesis that drives experimental determination of intron-exon structures. Therefore, it is clear that with the accumulation of genomic data from other species and a better understanding of the mechanisms and the signals involved in the transfer of information from sequence to function, more accurate computational models will be available. Those models have to face not only the complexity inherent to the biological processes and their regulatory pathways, but also the complexity of the inter- and intra-specific variability due to evolutionary events that led to the actual genomes of individuals and populations.

Existing gene finding programs, although significantly advanced over those that were available a few years ago, still have several important limitations. Almost without exception, computational gene finders predict only the coding fraction of a single spliced form of non-overlapping, canonical protein-coding genes. Annotation pipelines are currently able to extend those annotations by incorporating other biological features of clear interest for the research community, including non-coding mRNAs, pseudogenes, regulatory elements and transcription start sites, anti-sense transcripts, but also other genome-scale data collections such as gene expression profiles, protein interaction and genetic variation. However, a better understanding of the molecular mechanisms involved in gene expression and the integration of this knowledge into the theoretical models underlying the gene prediction software, may lead to systems that will be accurate enough to render both experimental verification and manual curation largely unnecessary [Brent and Guigó, 2004]. As more animal genomes are sequenced, deeper sequence alignments will contribute further to the definition of signals such as regulatory elements. The application of comparative genomics to study gene regulation has focused largely on the identification of shared regulatory sequences to explain similar patterns of gene expression between species. By contrast, the differences in gene regulation between organisms, and the role of these differences in speciation, have only just begun to be examined [Pennacchio and Rubin, 2001].

As more evidence of the conservation of exonic structures between orthologous genes and the sequence features that define such exons are accumulated [Waterston *et al.*, 2002; Gibbs *et al.*, 2004; Hillier *et al.*, 2004; Abril *et al.*, 2005], the analysis of the extent of that conservation becomes relevant to the prediction of alternative splicing events. Further evidence suggests that a large fraction of alternative splicing events is conserved between related species, such as human and mouse [Thanaraj *et al.*, 2003]. The analysis of the conserved sequence features involved in splice site definition, as well as in the regulation of splicing, will shed light on the code that determines the final pool of eukaryotic genes products. Alternative splicing remains, however, as a poorly solved problem. On the other hand, a comparison of the structural and mechanistic features of the major-class and minor-class, U2- and U12-types respectively, spliceosomes has provided many valuable insights into the essential catalytic elements of the splicing reaction. The rate-limiting excision of U12-type introns and their use in alternative expression of proteins *in vivo* indicates that they might be potential targets of gene regulation. Assessing gene expression patterns in transgenic organisms with U12 to U2 intron mutations should provide vital evidence and help to rationalize the continued presence of these rare introns in metazoan genomes [Patel and Steitz, 2003]. The existence of a second spliceosome raises the possibility that a third or fourth might be awaiting discovery. The degeneracy of the consensus sequences defining those signals would make yet another class of introns difficult to detect. Indeed, the GT-AG U12-type introns might well have been ignored for the initial focus on AT-AC introns.

Another promising research area involves the analysis of the polymorphisms that fall within the sequences defining splice sites or in the splicing regulatory sequences. Mutations in exonic or intronic regulatory elements that cause severe splicing defects might just be the tip of the iceberg. There might be also many genomic variants, including small indels and single nucleotide polymorphisms (SNPs), that cause partial splicing defects that are only pathogenic in specific tissues under the influence of a set of specific regulatory splicing factors. Similar to splicing, all those processes are rarely considered when assessing the clinical significance of genomic variants [Pagani and Baralle, 2004]. In this regard, we have gathered a database, to be explored in future analyses, which integrates gene structures for reference human genes [REFSEQ; Pruitt *et al.* 2005], the conservation scores from phylo-HMM based multiple alignments (for human, chimpanzee, mouse, rat, and chicken, and downloaded from the UCSC GENOME BROWSER; Karolchik *et al.* 2003) and a large collection of human SNPs from NCBI DBSNP [Sherry *et al.*, 2001].

Visualization tools will continue to play a key role in the integration of the genomic annotation data sets, in order to extract biological meaning from that flood of information. Due to the intrinsic dynamic nature of the annotation data sets, database browsers have become standard tools at the laboratory to retrieve the latest updates on genomic annotations and to navigate through the many different databases available. All public genome browsers have their particular strengths: the UCSC GENOME BROWSER exemplifies speed; NCBI MAP VIEWER is integrated into a larger site and is linked to the impressive range of databases that NCBI curates; `GBrowse` is a sophisticated toolkit designed to simplify building data browsers to display custom data; ENSEMBL provides flexibility and a broad range of data displays [Stalker *et al.*, 2004]. Notwithstanding, command-line flexible visualization tools still have their niche, as it is the case for `gff2ps`, `gff2aplot`, `compi` and similar tools. Although raster graphics are more popular and are currently best supported by web browsers, we still advocate the use of vector graphics to visualize genomic annotations. Vector graphics have one feature that makes them invaluable for many applications: they can be scaled without loss of image quality. For a long time, POSTSCRIPT has been the *de facto* standard of the graphics industry, and it has been well supported on `*nix` systems which provided not only interpreters, such as `ghostscript`, but also graphical interfaces for those interpreters, such as `ghostview`. With the advent of `XML` technologies, an emerging new graphics standard, the Scalable Vector Graphics format (`SVG`) will become the successor of POSTSCRIPT, at least for distributing vector graphics on the Internet. However, POSTSCRIPT is by itself a programming language. When self-contained documents are created, the data and the code to visualize such data share a single file, as happens for instance with `gff2ps` output.

In conclusion, finding all functional elements of genome sequences and using this information to improve the health of individuals and society, are the focus of the next phase of the Human Genome Project [Collins *et al.*, 2003]. Comparative analyses from multiple species at varying evolutionary distances are a powerful approach for identifying coding and functional non-coding sequences, as well as sequences that are unique for a given organism. Those techniques will continue to play a major role in the accurate annotation procedures required to understand the puzzling patchworks that are our genomes.

Chapter 7

Conclusions

> Errors, like straws, upon surface flow;
> he who would search for pearls must dive below...
>
> —John Dryden, "*All for love*"

In short, the research presented here has contributed to:

1. The development of a semi-automatic computational pipeline to perform whole genome analyses when comparing the human and mouse genomes. The main results are described hereunder:

   (a) The analyses included the production of gene predictions by `geneid`, an "*ab initio*" gene-finding software, and `SGP2`, initially a wrapper for `TBLASTX` and `geneid` to perform pair-wise comparative gene-finding.

   (b) Moreover, the evaluation of the predictions using a reference set of annotations and the visualization of the results, were among the steps of this pipeline.

   (c) The results from this pipeline, together with those provided by the people from the `Twinscan` project, were filtered by Genís Parra. Using an enrichment protocol based on the conservation of exonic structure between orthologous predictions between human and mouse, he supplied gene candidates for RT-PCR amplification to validate such predictions.

   (d) Several programs from this analysis pipeline have been adapted by Francisco Câmara. Currently, they are routinely used to predict genes on each new assembly version of several eukaryotic genomes. These species include human, mouse, rat, chicken, fruitfly, and the list keeps growing.

2. Describing the signals delimiting the boundaries between exons and introns. Taking advantage of the conservation of the exonic structures of orthologous genes in vertebrates, we have been able to tackle the comparative analysis of splice sites from orthologous introns. This research yielded the following results:

   (a) Human introns are on average larger than their respective orthologs in rodents. This can be explained by an increase in the repetitive sequences within those introns in the human lineage or by a loss of such repeats in the rodents lineage. The analysis of the distribution of ancient repeats, predating the split between human and rodents, supports the latter.

(b) We provide insights into the dynamics of the evolution of splice site sequences within four vertebrate genomes: human, mouse, rat and chicken. Our results confirm that the splicing code is under evolution, albeit very slow, remaining largely homogeneous within tetrapoda and showing noticeable differences only at larger phylogenetic distances.

(c) The greater conservation observed in mammalian/chicken orthologous splice sites compared to unrelated sites indicates that nucleotide substitution since the mammalian/avian split has not yet reached saturation at these sites. Saturation has been reached at intronic sites, which show a conservation level similar to that of unrelated sequences.

(d) The characteristic conservation of orthologous splice sites suggests that comparative prediction of splicing could improve methods based on the analysis of a single genome. Comparative prediction of splice sites could be particularly relevant to the prediction of alternative splicing features, a problem far from being solved.

(e) Our results seem to indicate that `U2` and `U12` introns have evolved independently after the split of mammals and birds, since we have not been able to document a single convincing case of conversion between these two types of introns.

(f) Furthermore, comparison of orthologous introns has also allowed us to define better the sequences involved in the specification of `U12` introns. These sequences, while more conserved than signals involved in `U2` intron specification, are more degenerate than previously thought.

3. The implementation of visualization tools for annotations obtained by gene-finding tools on genomic sequences, such as `gff2ps`, and to summarize the outcomes of comparative analyses, such as `gff2aplot` and `compi`. The main results are listed below:

   (a) `gff2ps` was devised to provide scalability and a flexible customization of the annotation feature attributes.

   (b) We have applied `gff2ps` to the “cartography” of sequence features for whole genomes of human, the fruitfly and the malaria mosquito. In those cases we had to implement specific software to integrate the large annotation data sets from these genomes and to provide specific customization parameters.

   (c) `gff2aplot` produces pair-wise alignment plots along with the annotation features of the sequences.

   (d) `compi` extends pictograms to compare the nucleotide frequencies of sequence patterns side-by-side. We used this tool in our orthologous splice sites signal comparison.

   (e) Several of the tools we have developed, including `gff2ps` and `gff2aplot`, have been made publicly available at our web site. They have been used with success by other groups to visualize the results of their own research.

APPENDICES

There and back again...
—Bilbo Baggins, "*The Hobbit*"

Curriculum Vitae

Josep F. Abril graduated on 1998 in Biology (Bachelor's degree) by **Universitat de Barcelona** (UB). He spent his last years as undergraduate collaborating with the laboratory of genome analysis at the **Grup de Recerca en Informàtica Biomèdica** (GRIB), under Dr. Roderic Guigó supervision. He obtained the Research competence and Advanced Studies Diploma ("*Diploma d'Estudis Avançats*", DEA) on 2002 by **Department of Experimental Sciences and Health** of **Universitat Pompeu Fabra** (UPF). From late 1998 till early 2005 he stayed as a *PhD* student, under supervision of Dr. Roderic Guigó within the GRIB.

Since 2000, he has been in charge of the **Genome Bioinformatics Lab** web site[1]. Among different software developments, it is worth to mention his contributions to visualization of genomic annotations, `gff2ps` and `gff2aplot`. `gff2ps` was used to visualize different whole genome maps, including those for human, the fruitfly and the malaria mosquito.

He has been teaching assistant for the practicals of the Bioinformatics course taught by the the **Genome Bioinformatics Laboratory** at **Universitat Pompeu Fabra**, between 2001 and 2005. Additionally, he has taught an introductory `perl` course for the **MSc on Bioinformatics for Health Sciences**, co-directed by **Universitat Pompeu Fabra** and **Universitat de Barcelona**, in 2004. He participated in the organization and presentation of the Bioinformatics stand for the "*Fira Viu la Ciència Contemporània*" (FVCC'03, a popular science fair) organized by the **Societat Catalana de Biologia**, in Barcelona in May 2003. He was also one of the organizers and lecturers of the workshops on "Computational Analysis of DNA Sequences" by **La Caixa**, held in Barcelona in November 2003 and 2004, and in Madrid in June 2004. He was invited speaker on the 4*th* meeting of the **Sociedad Española de Genética** held in El Escorial on October 2003.

He is currently involved in the management of paper contributions to the **4*th* European Conference on Computational Biology** (ECCB'05[2]), to be held in Madrid, Spain (September 28–October 1, 2005). He is also participating in the organization of the **ENCODE Genome Annotation Assessment Project** (EGASP'05[3]) workshop, to be held at the **Wellcome Trust Sanger Institute** (May 6–7, 2005).

His main research interest focuses on the computational analysis of the exonic structure of eukaryotic genes, its definition, evolution and association with genetic disorders. Genefinding and the visualization of genomic annotations are also within those interests.

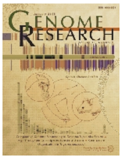
[1] GBL @ GRIB[IMIM-UPF-CRG] at: `http://genome.imim.es/`

[2] ECCB'05 at: `http://www.eccb05.org/`

[3] EGASP'05 at: `http://genome.imim.es/gencode/workshop2005.html`

List of Publications

Articles



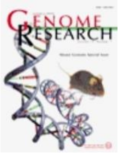
J.F. Abril, R. Castelo and R. Guigó.
“Comparison of splice sites in mammals and chicken.”
Genome Research, 15(1):111–119, 2005.

International Chicken Genome Sequencing Consortium (including J.F. Abril).
“Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution.”
Nature, 432(7018):695–716, 2004.

Rat Genome Sequencing Project Consortium (including J.F. Abril).
“Genome sequence of the brown Norway rat yields insights into mammalian evolution.”
Nature, 428(6982):493–521, 2004.

J.F. Abril, R. Guigó and T. Wiehe.
“`gff2aplot`: Plotting sequence comparisons.”
Bioinformatics, 19(18):2477–2479, 2003.

R. Guigó, E.T. Dermitzakis, P. Agarwal, C.P. Ponting, G. Parra, A. Reymond, J.F. Abril, E. Keibler, R. Lyle, C. Ucla, S.E. Antonarakis and M.R. Brent.
“Comparison of mouse and human genomes followed by experimental verification yields an estimated 1,019 additional genes.”
Proc. Nat. Acad. Sci., 100(3):1140–1145, 2003.



G. Parra, P. Agarwal, J.F. Abril, T. Wiehe, J.W. Fickett and R. Guigó.
“Comparative gene prediction in human and mouse.”
Genome Research, 13(1):108–117, 2003.

Mouse Genome Sequencing Consortium (including J.F. Abril).
"Initial sequencing and comparative analysis of the mouse genome."
Nature, 420(6915):520–562, 2002

R.A. Holt *et al* (including J.F. Abril).
"The Genome Sequence of the Malaria Mosquito *Anopheles gambiae*."
Science, 298(5591):129–149, 2002.

G. Glökner, L. Eichinger, K. Szafranski, J.A. Pachebat, A.T. Bankier, P.H. Dear, R. Lehmann, C. Baumgart, G. Parra, J.F. Abril, R. Guigó, K. Kumpf, B. Tunggal, the *Dictyostelium* Genome Sequencing Consortium, E. Cox, M.A. Quail, M. Platzer, A. Rosenthal and A.A. Noegel.
"Sequence and Analysis of Chromosome 2 of *Dictyostelium discoideum*."
Nature, 418(6893):79–85, 2002.

J.C. Venter *et al* (including J.F. Abril).
"The Sequence of the Human Genome."
Science, 291(5507):1304–1351, 2001.

T. Thomson, J.J. Lozano, R. Carrió, F. Serras, N. Loukili, M. Valeri, B. Cormand, M.P. del Río, J.F. Abril, M. Burset, E. Sancho, J. Merino, A. Macaya, M. Corominas and R. Guigó.
"Fusion of the human gene for the polyubiquitination co-effector uev-1 with kua, a newly identified gene."
Genome Research, 10(11):1743–1756, 2000.

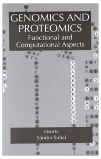
J.F. Abril and R. Guigó.
"`gff2ps`: visualizing genomic annotations."
Bioinformatics, 16(8):743–744, 2000.

R. Guigó, P. Agarwal, J.F. Abril, M. Burset and J.W. Fickett.
"An Assessment of Gene Prediction Accuracy in Large DNA Sequences."
Genome Research, 10(10):1631–1642, 2000.

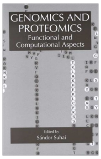
M.G. Reese, G. Hartzell, N.L. Harris, U. Ohler, J.F. Abril and S.E. Lewis.
"Genome Annotation Assesment in *Drosophila melanogaster*."
Genome Research, 10(4):483–501, 2000.

M.D. Adams *et al* (including J.F. Abril).
"The Genome Sequence of *Drosophila melanogaster*."
Science, 287(5461):2185–2195, 2000.

Book Chapters



J.F. Abril, S. Castellano and R. Guigó.
"Comparative gene prediction."
In M.D. Adams editor:
Comparative Genomics: A Guide to the Analysis of Eukaryotic Genomes.
Humana Press, 2004 (*in press*).



R. Guigó, M. Burset, P. Agarwal, J.F. Abril, R.F. Smith and J.W. Fickett.
"Sequence Similarity Based Gene Prediction."
In S. Suhai editor:
Genomics and Proteomics: Functional and Computational Aspects.
Plenum Publishing Corporation, 2000. ISBN: 0–306–46312–1.

Posters

J. Lagarde, J.F. Abril, F. Denoeud, R. Guigó and the GENCODE Consortium.
"`ENr334`: Computational Gene Predictions, VEGA Annotations and GENCODE Experimental Validations."
CSHL - Genomics Workshop "Identification of Functional Elements in Mammalian Genomes", New York, USA (2004)

J.F. Abril, M. Albà, E. Blanco, M. Burset, F. Câmara, S. Castellano, R. Castelo, O. Gonzalez, G. Parra and R. Guigó.
"Understanding the Eukaryotic Genome Sequence."
Inaugural Symposium of the Center for Genomic Regulation, Barcelona, Spain (2002)

E. Blanco, G. Parra, S. Castellano, J.F. Abril, M. Burset, X. Fustero, X. Messeguer and R. Guigó.
"Gene Prediction in the Post-Genomic Era."
IXth ISMB, Copenhagen, Denmark (2001)

G. Glöckner, L. Eichinger, K. Szafranski, P. Dear, J. Pachebat, K. Kumpf, R. Lehmann, J.F. Abril, G. Parra, R. Guigó, B. Tunggal, E. Cox, M.A. Quail, M. Platzer, A. Rosenthal, A.A. Noegel and the *Dictyostelium* Genome Sequencing Consortium.
"Sequence and Analysis of Chromosome 2 from the Model Organism *Dictyostelium discoideum*."
CSHL - Genome Sequencing & Biology, New York, USA (2001)

J.F. Abril, E. Blanco, M. Burset, S. Castellano, X. Fustero, G. Parra and R. Guigó.
"Genome Informatics Research Laboratory: Main Research Topics."
Ist Jornadas de Bioinformática, Cartagena, Spain (2000)

T. Wiehe, J.F. Abril, M. Burset, S. Gebauer-Jung and R. Guigó.
"Comparative Genomics: At the Crossroads of Evolutionary Biology and Genome Sequence Analysis."
VIIth ESEB, Barcelona, Spain (1999)

T. Wiehe, J.F. Abril, M. Burset, S. Gebauer-Jung and R. Guigó.
"Gene Prediction and Validation Based on Homologous Genomic Sequences."
VIIth ISMB, Heidelberg, Germany (1999)

J.F. Abril, T. Wiehe, M. Burset and R. Guigó.
"Tools to Visualize Genome Annotations."
IIIrd RECOMB, Lyon, France (1999)

M. Burset, J.F. Abril and R. Guigó.
"`GeneID-3`, from DNA Sequence to Protein Function."
Vth ISMB, Halkidiki, Greece (1997)

Contact Information

Find below, in alphabetical order, the contact information of some of the authors of the research presented here:

Josep F. Abril Ferrando — *PhD Researcher*
Genome Bioinformatics Research Lab
Research Group in Biomedical Informatics
Institut Municipal d'Investigació Mèdica
Dr. Aiguader, 80, 08003 Barcelona (Spain)
Phone: +34 93 224 0890 || Fax: +34 93 224 0875
E-mail: `jabril` *at* `imim.es`
Web: http://genome.imim.es/~jabril/

Mark D. Adams — *Associate Professor*
Department of Genetics
Case Western Reserve University
10900 Euclid Avenue, Cleveland, OH 44106 (USA)
Phone: +01 216 368 2791
E-mail: `mda13` *at* `cwru.edu`
Web: http://genomics.case.edu/people_adams.html

Pankaj Agarwal — *Investigator*
Department of Bioinformatics
GlaxoSmithKline Pharmaceuticals R&D
709 Swedeland Road, UW2230, King of Prussia, PA 19406-0939 (USA)
E-mail: `pankaj.agarwal` *at* `gsk.com`

Ewan Birney — *Research Group Leader*
EMBL Outstation - Hinxton
European Bioinformatics Institute
Wellcome Trust Genome Campus
Hinxton, Cambridge, CB10 1SD (United Kingdom)
Phone: +44 (0)1223 494 420 || Fax: +44 (0)1223 494 468
E-mail: `birney` *at* `ebi.ac.uk`
Web: http://www.ebi.ac.uk/~birney/

Robert Castelo Valdueza — *Senior Researcher*
Genome Bioinformatics Research Lab

Research Group in Biomedical Informatics
Institut Municipal d'Investigació Mèdica
Dr. Aiguader, 80, 08003 Barcelona (Spain)
Phone: +34 93 224 0884 || Fax: +34 93 224 0875
E-mail: rcastelo *at* imim.es
Web: http://genome.imim.es/~rcastelo/

Roderic Guigó i Serra — *Research Group Leader*
Genome Bioinformatics Research Lab
Research Group in Biomedical Informatics
Institut Municipal d'Investigació Mèdica
Dr. Aiguader, 80, 08003 Barcelona (Spain)
Phone: +34 93 224 0877 || Fax: +34 93 224 0875
E-mail: rguigo *at* imim.es
Web: http://genome.imim.es/~rguigo/

Robert Holt — *Research Head*
Sequencing Group
Genome Sciences Centre
BC Cancer Research Centre
Suite 100, 570 West 7th Ave, Vancouver, BC, V5Z 4S6 (Canada)
Phone: +01 604 877 6276
Email: rholt *at* bcgsc.ca
Web: http://www.bcgsc.ca/about/faculty/person?pid=rholt

Genís Parra Farré — *PhD Researcher*
Genome Bioinformatics Research Lab
Research Group in Biomedical Informatics
Institut Municipal d'Investigació Mèdica
Dr. Aiguader, 80, 08003 Barcelona (Spain)
Phone: +34 93 224 0884 || Fax: +34 93 224 0875
E-mail: gparra *at* imim.es
Web: http://genome.imim.es/~gparra/

Martin G. Reese — *Investigator*
Omicia Inc.
5980 Horton Street, Suite 235, Emeryville, CA 94608 (USA)
Phone: +01 510 595 0800 || Fax: +01 510 588 4523
E-mail: mreese *at* omicia.com

Thomas Wiehe — *Research Group Leader*
Institut fuer Molekularbiologie und Biochemie
Freie Universität Berlin
Berlin Center for Genome Based Bioinformatics
Arnimallee 22, 14195 Berlin (Germany)
Phone: +49 30 8445 1504 || Fax: +49 30 8445 1504
E-mail: twiehe *at* zedat.fu-berlin.de
Web: http://www.bcbio.de/jrg_wiehe/

Miscellanea

This thesis layout is largely derived from the LaTeX template created by Robert Castelo in 2002[1]. His templates were extended by Sergi Castellano and Genís Parra for their theses (see the corresponding references in page 254). The templates on which this document was built were derived from them. Here, some comments on it and the source code for download are provided.

Technical comments

This book was typeset with GNU `emacs` 21.3.1 in LaTeX mode and converted to PDF with `pdflatex` 3.14159-1.10b (Web2C 7.4.5). All running on a linux box with Red Hat Fedora Core 2 and kernel 2.6.9-1.6. LaTeX is a document preparation system, powerful, robust and able to achieve professional results [Lamport, 1994]. However, the learning curve may be stiff. Therefore, a link to an initial template is given at the end of this chapter for your convenience.

The main document, `thesis.tex`, depends on several LaTeX files—including each chapter, the tables and few POSTSCRIPT figures—, but it also depends on other files—such as style files, hacked LaTeX packages, several bitmaps and the PDF files for the attached papers—. Furthermore, `pdflatex` had to be run several times, together with BIBTeX (to produce the bibliography chapter), `makeindex` (to build the index, the glossaries and the acronyms list), `thumbpdf` (to generate the main PDF document thumbnails), and few `perl` scripts. A `Makefile` was written to automatize the compilation process of the whole document. In fact, the `Makefile` was extended to produce four versions of the main document. The "*draft*" version does not include figures and the PDF files for the papers, and it displays crop marks and boxes around several elements (such as the area reserved for the pictures). The "*proofs*", where everything is included but crop marks and boxes are kept, and different hyperlink types use different colors. The "*pdf*" version is the electronic version in which all the hyperlinks are marked in blue color, crop marks are disabled. Finally, the "*press*" version is very similar to the "*pdf*" one, currently the only difference is that all the hyperlinks are black (to save some money when printing the hardcopy, of course). The `Makefile` also includes a rule to build the final book "*cover*", which recycles

[1] R. Castelo, April 2002.
"The Discrete Acyclic Digraph Markov Model in Data Mining"
Faculteit Wiskunde en Informatica, Universiteit Utrecht

the `abstract.tex` file and takes some customization from the same style file as the main `thesis.tex` file.

The compilation of a complete version of this document takes about 600 seconds—of course, the "*draft*" version takes much less—with an AMD Athlon 64 processor 3200+, with 512 KB of RAM. This is mainly due to the several steps required to ensure that every reference, index and so on, is in place. The basic build series of commands is the following: an initial `pdflatex`, a BIBTEX run to produce the bibliography, a second run of `pdflatex` to include it, three calls to `makeindex` (one for the Acronyms Glossary, another for the Web Glossary and the last for the standard Glossary of terms), a third run of `pdflatex` to include the glossaries, another call to `makeindex` (to generate the final index) and to `pdflatex`, then `makeindex` and `pdflatex` are run again, an extra run of `pdflatex` is followed by `thumbpdf`, and a final `pdflatex` to obtain the finished document. If any problem was found, like missing references, an extra round of `pdflatex`, BIBTEX and `pdflatex` is performed by the `Makefile`.

Here you can find the version of some of the programs refereed above: BIBTEX version 0.99c (Web2C 7.4.5), `thumbpdf` version 3.2 (2002/05/26), and `makeindex` version 2.14 (2002/10/02).

LATEX Packages

As there are four versions of the document, the `ifthen` package was used to define version specific parameters, as well as to include different files. The package `geometry` facilitates the definition of the page layout. The current document original dimensions for both, the electronic and printed versions, are 170 mm width by 240 mm height. The "*cover*" requires `calc` to calculate automatically the total width for the page layout, which includes the front and the back covers and the spine width. The main document basic font size is the default value for the "`book`" document class, 10 pt.

The `crop` package is usefull to define the trimming marks for the "*draft*" and "*proofs*" versions of this document. It distinguishes between the logical page, the page sizes defined by the user, and the physical page, the page size for the hardcopy. The `layout` package is used in the "*draft*" version to show on the first page the LATEX variable settings controlling the page layout. Another useful package has been `nextpage`, which provides additional "`clear...page`" commands that ensure to get empty even pages at the end of chapters—and of course, to ensure that all chapters begin at odd pages—, even with automatically generated sections like the Bibliography and the Index.

The `babel` package provides a set of options that allow the user to choose the language(s) in which the document will be typeset, for instance language-specific hyphenation patterns. The default language was set to "`english`", while "`catalan`" and "`spanish`" were also loaded for using them for the corresponding translations of the ABSTRACT (see pages xxv and xxvii respectively).

When working with `pdflatex` there are three unvaluable packages: `pdfpages`, which makes it easy to embed external PDF documents, such as the attached publications (see for instance page 158); `thumbpdf`, it must be included in files for which a user wants to generate thumbnails (which are created by the `thumbpdf` program); and `hyperref`, which extends the functionality of all the LATEX cross-referencing commands to produce `special` commands which a driver can turn into hypertext links. To protect URL characters we must load the `url` package, unless we have already provided `hyperref`. This package has its own version of the `url` macro, enhanced to provide clickable URLs.

To include POSTSCRIPT figures one needs `graphics` and/or `graphicx`, those packages are modified by **`pdflatex`** so that they are able to include bitmaps (PNGs, JPEGs, and so on) and PDF files into the document. `color` facilitates the specification of user-defined colors (such as the cover green shades). Figures generated with LATEX can use any of the following packages: `pstricks`, `pstcol`, `multido`.

The bibliography was produced with BIBTEX. The package `natbib` (NATural sciences BIBliography) provides both author-year and numerical citations; and it makes possible to define different citation styles. We have set the following options: "`square`", to put citations within square brackets; "`colon`", to separate multiple citations with colons; and "`authoryear`" to show author and year citations (instead of numerical citations). The style "`plainnat`" was then applied to format the bibliography.

`makeidx` provides the macros required to make a subject index. To show the capital letter section headings, few variables were redefined on an auxiliary file (`header.ist`). Three glossaries were generated for this document: the acronyms (see page 203), the web references (see page 213) and the glossary of terms (see page 207). The package `glossary` allowed us to customize the format of these three sections.

We also defined a style file named `mythesis.sty`. It loads the following font packages: `fontenc` (with "`T1`" option), to set extended font encoding (accents and so on); `textcomp`, to include some extra symbols, such as the Euro symbol for instance; `pifont`, for SYMBOL and ZAPF DINGBATS fonts; `mathpazo`, with which roman family and formulas are set to PALATINO; `avant`, with which sans-serif family is set to AVANT GARDE; and `courier`, to set typewriter family to COURIER. Accessory documents, such as LATEX-generated figures, can use the following font packages: `times`, `t1enc`, and `helvet`.

Other packages that were loaded are: `fancyhdr`, to produce nice headings; `fancyvrb`, to extend the `verbatim` environment; `comment`, to hide parts of the original LATEX files; `rotating`, to rotate boxes of text; and `multirow`, to get multirow cells within the `tabular` environment.

Getting the template files

The template for this document and all related files will be available from:

`http://genome.imim.es/~jabril/thesis/`

Abbreviations

3′ss	3′ Splice Site (intronic, acceptor site)
5′ss	5′ Splice Site (intronic, donor site)
aa	Amino Acids (protein sequence length unit)
ACT	Artemis Comparison Tool
ASD	Alternative Splicing Database
BLAST	Basic Local Alignment Search Tool
BLAT	BLAST-Like Alignment Tool
bp	Base Pairs (nucleotide sequence length unit)
CDS	CoDing Sequence (protein-coding)
CTD	Carboxy-Terminal Domain (of RNApolII)
DAS	Distributed Annotation System
DNA	DeoxyriboNucleic Acid
EBI	European Bioinformatics Institute
ECR	Evolutionary Conserved Regions
EHMM	Evolutionary Hidden Markov Model
EJC	Exon-Junction Complex
ENCODE	ENCyclopedia Of DNA Elements
ESE	Exonic Splicing Enhancer
ESS	Exonic Splicing Silencer
FMM	First-order Markov Model
FTP	File Transfer Protocol

GASP	Genome Annotation Assessment Project
GFF	General Feature Format
GHMM	Generalized Hidden Markov Model
GNU-GPL	GNU General Public License
GPHMM	Generalized Pair HMM
HAVANA	Human And Vertebrate Analysis aNd Annotation
HMM	Hidden Markov Model
ICGSC	International Chicken Genome Sequencing Consortium
IHGSC	International Human Genome Sequencing Consortium
IMGSC	International Mouse Genome Sequencing Consortium
ISE	Intronic Splicing Enhancer
ISS	Intronic Splicing Silencer
mRNA	Messenger RNA
mRNP	mRNA-protein Particle
NCBI	National Center for Biotechnology Information
ncRNA	Non-Coding RNA
NIH	National Institutes of Health
NISC	NIH Intramural Sequencing Center
NMD	Nonsense-Mediated mRNA Decay
ORF	Open Reading Frame
PHMM	Pair Hidden Markov Model
phylo-HMM	Phylogenetic Hidden Markov Model
PiPs	Percentage Identity Plots
PSM	Position-specific Scoring Matrix
PTC	Premature Termination Codon
PWM	Position Weight Matrix
RGSPC	Rat Genome Sequencing Project Consortium
RNA	RiboNucleic Acid
rRNA	Ribosomal RNA

Symbol	Meaning	Origin of designation
A	A	**A**denine
C	C	**C**ytosine
G	G	**G**uanine
T	T	**T**hymine
U	U	**U**racil
R	A or G	pu**R**ine
Y	C or T	p**Y**rimidine
M	A or C	a**M**ino
K	G or T	**K**etone
W	A or T	**W**eak interaction (2 H bonds)
S	C or G	**S**trong interaction (2 H bonds)
B	C or G or T	not-A, **B** follows A in the alphabet
D	A or G or T	not-C, **D** follows C
H	A or C or T	not-G, **H** follows G
V	A or C or G	not-T (not-U), **V** follows U
N	G or A or T or C	a**N**y (unspecified)
X	G or A or T or C	a**N**y (often meaning unknown)

Table E.1: **Extended DNA / RNA alphabet.** It includes symbols coding for nucleotide ambiguity. Adapted from IUPAC-IUB for nucleotide nomenclature [Cornish-Bowden, 1985].

SNP	Single Nucleotide Polymorphism
snRNP	Small Nuclear RiboNucleoprotein Particle
SVG	Scalable Vector Graphics
tRNA	Transfer RNA
U2AF	U2 Auxiliary Factor
UCSC	University of California, Santa Cruz
URL	Uniform Resource Locator
UTR	UnTRanslated sequence
VEGA	VErtebrate Genome Annotation
VRML	Virtual Reality Modeling Language
WABA	Wobble Aware Bulk Aligner

Symbols		Amino Acid	Codons
A	Ala	Alanine	**GCA** GCC **GCG** GCU
C	Cys	Cysteine	**UGC** UGU
D | B	Asp	Aspartic acid	**GAC** GAU
E | Z	Glu	Glutamic acid	**GAA** GAG
F	Phe	Phenylalanine	**UUC** UUU
G	Gly	Glycine	**GGA** GGC **GGG** GGU
H	His	Histidine	**CAC** CAU
I	Ile	Isoleucine	**AUA** AUC **AUU**
K	Lys	Lysine	**AAA** AAG
L	Leu	Leucine	**UUA** UUG **CUA** CUC **CUG** CUU
M	Met	Metionine	**AUG**
N | B	Asn	Asparagine	**AAC** AAU
P	Pro	Proline	**CCA** CCC **CCG** CCU
Q | Z	Gln	Glutamine	**CAA** CAG
R	Arg	Arginine	**AGA** AGG **CGA** CGC **CGG** CGU
S	Ser	Serine	**AGC** AGU **UCA** UCC **UCG** UCU
T	Thr	Threonine	**ACA** ACC **ACG** ACU
V	Val	Valine	**GUA** GUC **GUG** GUU
W	Trp	Tryptophan	**UGG**
Y	Tyr	Tyrosine	**UAC** UAU
X	Any	Unknown aa	**NNN**
*	(!)	Stop codon: ocre	**UAA**
*	(#)	Stop codon: amber	**UAG**
*	(@)	Stop codon: opal	**UGA**
U	Sec	Selenocysteine	**UGA**

Table E.2: **The standard genetic code.** Synonymous codons are alternatively boldfaced to ease their distinction. Single letter notation follows IUPAC-IUB for amino acid symbols [IUPAC-IUB JCBN, 1984, 1993]. Termination codons are listed separately and their extended symbol codes are shown in brackets. This extended notation was devised in our laboratory to distinguish each stop codon on translated sequences; i.e., when analyzing those sequences to look for selenocysteine amino acid codon corresponding to UGA termination codon [Hatfield and Gladyshev, 2002].

Glossary

Acceptor Splice Site

The binding site of the spliceosome on the 3′ side of an intron and the 5′ side of an exon. This term is preferred over 3′ site because there can be multiple acceptor sites, in which case 3′ site is ambiguous. Also, one would have to refer to the 3′ site on the 5′ side of an exon, which is confusing. Mechanistically, an acceptor site defines the beginning of the exon, not the other way around.

Algorithm

A systematic procedure for solving a problem in a finite number of steps, typically involving a repetition of operations. Once specified, an algorithm can be written in a computer language and run as a program. Named after an Iranian mathematician, Al-Khawarizmi.

Alignment

The procedure of comparing two or more sequences by looking for a series of individual characters or character patterns that are in the same order in the sequences. There are two type of alignments: local, which attempts to align regions of sequences with the highest density of matches (one or more islands of subalignments are created in doing so); and global, which attempts to match as many characters as possible, from end to end, in the set of sequences.

Annotation

The elucidation and description of biologically relevant features in the sequence is essential in order for genome data to be useful. The quality with which annotation is done will have direct impact on the value of the sequence. At a minimum, the data must be annotated to indicate the existence of gene coding regions and control regions. Further annotation activities that add value to a genome include finding simple and complex repeats, characterizing the organization of promoters and gene families, the distribution of G+C content, tying together evidence for functional motifs and homologs and so forth.

Capping

The process by which eukaryotic mRNA is modified by the addition at the 5′ terminus of an m7G(5′)ppp(5′)N structure. Capping is essential for several important steps of gene expression, for instance, mRNA stabilization, splicing, mRNA export from the nucleus and initiation of translation.

Consensus Sequence (consensus)

The simplest form of a consensus sequence is created by picking the most frequent base at some position in a set of aligned DNA, RNA or protein sequences. The process of creating a consensus destroys the frequency information and leads to many errors in interpreting sequences. It is one of the worst pitfalls in molecular biology. Suppose a position in a binding site had 75% A. The consensus would be A. Later, after having forgotten the origin of the consensus while trying to make a prediction, one would be wrong 25% of the time.

Conserved

Derived from a common ancestor and retained in contemporary related species. Conserved features may or may not be under selection.

Conserved Segments

Also known as **Conserved Linkages**, is a special case of the conserved synteny in which the order of multiple orthologous genes is the same in the compared species.

Distributed Annotation System

The distributed annotation system [DAS, Dowell *et al*. 2001] is a client-server system in which a single client integrates information from multiple servers. It allows a single machine to gather up genome annotation information from multiple distant web sites, collate the information, and display it to the user in a single view. Little coordination is needed among the various information providers.

Donor Splice Site

The binding site of the spliceosome on the $5'$ side of an intron and the $3'$ side of an exon. This term is preferred over $5'$ site because there can be multiple donor sites, in which case $5'$ site is ambiguous. Also, one would have to refer to the $5'$ site on the $3'$ side of an exon, which is confusing. Mechanistically, a donor site defines the end of the exon, not the other way around.

Dot-Plot

A graphical representation of the regions of similarity between two sequences. The two sequences are placed on the axes of a rectangular matrix and (in the simplest forms of dotplot) wherever there is a similarity between the sequences a dot is placed on that matrix. A dot-plot gives an overview of all possible alignments between two sequences, where each diagonal corresponds to a possible (ungapped) alignment.

Enhancer

Control element that elevates the levels of transcription from a promoter, independent of orientation or distance. Those intronic and exonic *cis*-acting elements stimulating splicing and that are important for correct splice-site identification.

Eukaryote

Organisms with intracellular membranous organelles such as the nucleus and mitochondria.

Exon

The segment of a pre-mRNA that contains protein-coding sequence and/or the $5'$ or $3'$ untranslated sequences, which must be spliced together with other exons to produce a mature mRNA.

Exon-definition model

A model in which exon units, rather than intron units, are initially defined by pairings of spliceosomal components across exons.

Gene

A functional unit of the genome. When not specifically stated, "gene" is usually considered a "protein-coding" gene, but many genes do not contain the instructions for proteins (see non-coding RNA).

Genome

The complete genetic material for an organism. All the DNA contained in an organism or a cell, which includes both the chromosomes within the nucleus and the DNA in mitochondria.

Genome Browser

A web-based or standalone software that serves as a front-end to navigate through a database of genomic annotations for one or more species. A genome browser stacks annotation tracks beneath genome coordinate positions, allowing rapid visual correlation of different types of information. The genome browser itself does not draw conclusions; rather, it collates all relevant information in one location, leaving the exploration and interpretation to the researcher.

Hidden Markov models

Probability models that were first developed in the speech-recognition field and later applied to protein- and DNA-sequence pattern recognition. Hidden Markov models (HMMs) represent a system as a set of discrete states and as transitions between those states. Each transition has an associated probability. Markov models are hidden when one or more of the states cannot be observed directly. HMMs are valuable in bioinformatics because they allow a search or alignment algorithm to be built on firm probability bases, and it is straightforward to train the parameters (transition probabilities) with known data.

Homologs

Features in species being compared that are similar because they are ancestrally related.

Homology Blocks

Also defined as **Conserved Synteny**, occurs when the orthologs of genes that are on the same chromosome in one species are also on the same chromosome in the comparison species.

Intron

An intervening non-coding sequence that interrupts two exons and that must be excised from pre-mRNA transcripts before translation.

Intron Branch Point

The adenosine residue near the 3′ end of an intron the 2′ hydroxyl group of which becomes linked to the 5′ end of the intron during the first step of splicing.

Intron-definition model

protect A model that proposes the initial pairwise interaction of spliceosomal components across introns, defining introns units that subsequently interact to promote spliceosome assembly and catalysis.

Lariat

An RNA, the 5′ end of which is joined by a phospodiester linkage to the 2′ hydroxyl of an internal nucleotide, thereby creating a lasso-shaped molecule.

Neural Networks

A collection of mathematical models that emulate some of the observed properties of biological nervous systems and draw on the analogies of adaptative biological learning. Many highly interconnected processing elements that are analogous to neurons, are tied together with weighted connections that are analogous to synapses. Once it is trained on known exon or intron sample sequences, it will be able to predict exons or introns in a query sequence automatically.

Non-Coding RNA

Some RNAs, like tRNAs or rRNAs, do not contain information for protein sequences. The RNA molecule for those genes defines a function by itself and does not need to get translated into protein.

Open Reading Frame

Each strand of DNA has three frames. Any subsequence that does not contain stop codons in a particular frame is an open reading frame.

Orthologs

Homologous features that separated because of a speciation event, they derive from the same gene in the last common ancestor. See Jensen [2001] for more information on this item.

Paralogs

Homologous features that separated because of duplication events.

Phylogenetic Distances

Measures of the degree of separation between two organisms or their genomes, expressed in various terms such as the number of accumulated sequence changes, number of years or number of generations. The distances are often placed on phylogenetic trees, which show the deduced relationships among the organisms.

Pip-Plot

Pip-plots display all the ungapped alignments between two sequences as black horizontal lines. The length of the line corresponds to the length of the alignment, while its height corresponds to the percent identity of the alignment. An example of a tool producing this output is `PipMaker` [Schwartz *et al.*, 2000].

Prokaryote

Organisms that do not contain intracellular membranous organelles. All bacteria are prokaryotes.

Promoter Element

A region of DNA extending 150-300 bp upstream from the transcription start site that contains binding sites for RNA polymerase and a number of proteins that regulate the rate of transcription of the adjacent gene. In RNA synthesis, promoters are a means to demarcate which genes should be used for messenger RNA creation—and, by extension, control which proteins the cell manufactures.

Proteome

The complete set of all proteins produced by a particular organism. Many proteins undergo post-translational modifications that add or subtract features from a protein. Therefore, a particular mRNA might have many different protein isoforms.

Pseudogene

A DNA sequence that was derived originally from a functional protein-coding gene that has lost its function, owing to the presence of one or more inactivating mutations.

Regulatory Element

A *cis*-acting DNA sequence that is required for a gene to be transcribed, or to be transcribed in the proper cell type(s) and developmental stage(s). These sequences are recognized by different transcription factors which modulate the binding or the activity of the RNA polimerase. These sequences comprise promoter regions, enhancers and

Sequence Pattern

A sequence pattern is defined by a set of aligned nucleotide or amino acid sequences (i.e. binding sites, splicing signals, and so on), or by a common protein structure. In contrast, consensus sequences, regular expressions, sequence logos and pictograms are only models of the patterns found experimentally or in nature. Models do not capture everything in nature. For example, there might be correlations between two different positions in a binding site. A more sophisticated model might capture these but still not capture three-way correlations. It is impossible to make the more detailed model if there is not enough data.

Silencer

Control element that supresses gene expression independent of orientation or distance. Those intronic and exonic *cis*-acting elements repressing splicing and that are important for correct splice-site identification.

Small Nuclear Ribonucleoprotein Particle

A particle that is found in the cell nucleus and consist of a tight complex between a short RNA molecule (up to 300 nucleotides) and one or more proteins. SnRNPs are involved in pre-mRNA processing and transfer RNA biogenesis.

Smooth-Plot

Smooth-plots are constructed using, for each nucleotide, a 100 bp sliding window in which sequence identity between two sequences is averaged. Such a window centered at every nucleotide in the base sequence is used to calculate the number of matches inside of this window. Percent identity counts in a sliding window are utilized to calculate the height of the smooth conservation graph at each point. Basically, smooth-graph is a smooth average of the Pip-plot. Smooth-graphs present a simplified and clearer view in the conservation profile but loses information regarding gap distribution in the alignment. An example of a tool producing this output is `VISTA` [Mayor *et al.*, 2000].

Spliceosome

A large complex that consist of five splicing small nuclear ribonucleoprotein particles as well as numerous protein factors. It mediates the excision of introns from pre-mRNA transcripts and ligates exon ends to produce mature mRNA.

Synteny

The property of being on the same chromosome *sensu strictu* [Passarge *et al.*, 1999]. Nowadays is often used as synonymous of **Homology Blocks**, specially within the gene-finding terminology.

Training Data Set

The known examples of an object (for example, an exon) that are used to train prediction algorithms, so that they learn the rules for predicting an object. They can be positive training sets (consisting of true objects, such as exons) or negative training sets (consisting of false objects, such as pseudogenes).

Transcriptome

The complete set of transcripts for a particular genome. This term is often used to mean the mRNAs of protein coding genes and their alternatively spliced variants.

WebSite References

AceDB genome database

AceDB is a genome database designed specifically for handling bioinformatic data flexibly. It includes tools designed to manipulate genomic data, but is increasingly also used for non-biological data.

`http://www.acedb.org/`

Analysis of mammalian and chicken splice sites

This web page summarizes the supplementary materials for Abril *et al.* [2005].

`http://genome.imim.es/datasets/hmrg2004/`

Assessment of gene prediction accuracy in large DNA sequences

Given the absence of experimentally verified large genomic data sets, a semi-artificial test set comprising a number of short single-gene genomic sequences with randomly generated intergenic regions was built in order to analyze gene-prediction programs accuracy [Guigó *et al.*, 2000].

`http://genome.imim.es/datasets/gpeval2000/`

`compi` home page

`compi` is a `perl` script to produce *comparative pictograms*, a graphical representation of nucleotide frequencies at each position of a sequence motif or a pair-wise comparison between two sequence patterns. Latest version, as well as examples, of this program will be available from the URL below:

`http://genome.imim.es/software/compi/`

Ensembl Genome Browser

Ensembl is a joint project between EMBL - EBI and the Sanger Institute to develop a software system which produces and maintains automatic annotation on metazoan genomes. The following URL corresponds to the project main page:

`http://www.ensembl.org/`

Gene Predictions on Genomes

A repository of gene predictions on eukaryotic genomes. It contains the results from geneid and SGP2 when applied on each novel genome assembly. Annotations for several species, including human, chimp, mouse, rat, chicken and the fruitfly, can be retrieved from:

http://genome.imim.es/genepredictions/

geneid predictions submitted to GASP1

A set of training sequences (exons/introns) and the resulting parameters required to run geneid on *Drosophila melanogaster* genome.

http://genome.imim.es/datasets/Dro_me/

General Feature Format (GFF)

Initially proposed at Sanger Center by Richard Durbin and David Haussler in 1997, it was proposed as a protocol for the transfer of annotation features information. It has undergone two major reviews, each one defining a new version (GFF v1, v2 and v3). It also inspired a derivated format known as Gene Transfer Format (GTF, http://genes.cs.wustl.edu/GTF2.html), which has additional structure that warrants a separate definition and format name. Main fields of the GFF format are:

```
seqname source feature start end score strand frame [attributes] [# comments]
```

Further information is available at:

http://www.sanger.ac.uk/Software/formats/GFF/GFF_Spec.shtml

Generic Model Organism Project

The home page of a joint effort by the model organism system databases WORMBASE, FLYBASE, MGI, SGD, GRAMENE, RAT GENOME DATABASE, ECOCYC, and TAIR to develop reusable components suitable for creating new community databases of biology.

http://www.gmod.org/

Genome Annotation Assessment Project (GASP1)

Community wide experiment to assess gene prediction on long eukaryotic genomic sequences: The Adh region (2.9Mb) in *Drosophila melanogaster*.

http://www.fruitfly.org/GASP1/

gff2aplot home page

gff2aplot is a tool for generating pair-wise alignment-plots for genomic sequences in POSTSCRIPT [Abril *et al.*, 2003]. Latest version of this program can be retrieved from this URL, as well as examples and tutorials on how to use it.

http://genome.imim.es/software/gfftools/GFF2APLOT.html

gff2ps home page

This is the home page for gff2ps, a program for visualizing annotations of genomic sequences [Abril and Guigó, 2000]. The program takes as input the annotated features on a genomic sequence in GFF format, and produces a visual output in POSTSCRIPT. It has been successfully used to generate the whole genome maps of different eukaryotic organisms, including human. Latest version of this program can be retrieved from this URL, as well as examples and tutorials on how to use it.

http://genome.imim.es/software/gfftools/GFF2PS.html

Making the three panels poster for the ISMB99 GASP1 tutorial

The posters made for the GASP1 tutorial and shown at ISMB'99 meeting are an example of what can be done with the gff2ps visualization tool. There you will find three examples of what can be generated from the same data-set, applying a slightly modified customization file and few command-line options.

http://genome.imim.es/software/gfftools/GFF2PS-ADHposter.html

Mouse genome supplementary materials

Description of the software and data presented in Guigó *et al.* [2003] and Waterston *et al.* [2002]. In that paper it was estimated that near a thousand novel human genes that do not overlap known proteins can be verified experimentally. The method is based in the comparison of human and mouse genomes to enhance the resulting gene-predictions, plus a filtering step from which a sample of mouse predictions were tested by RT-PCR amplification and direct sequencing.

http://genome.imim.es/datasets/mouse2002/

NCBI MAP VIEWER

The NCBI MAP VIEWER provides special browsing capabilities for a subset of organisms in ENTREZ Genomes (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Genome). Available organism genomes are listed on the NCBI MAP VIEWER Home Page. This browser allows the visitor to view and search an organism's complete genome, display chromosome maps, and zoom into progressively greater levels of detail, down to the sequence data for a region of interest.

http://www.ncbi.nlm.nih.gov/mapview/

RepeatMasker

RepeatMasker is a program that screens DNA sequences for interspersed repeats and low complexity DNA sequences. The output of the program is a detailed annotation of the repeats that are present in the query sequence as well as a modified version of the query sequence in which all the annotated repeats have been masked (by default replaced by Ns).

http://www.repeatmasker.org/

SGP2 home page

SGP2 is a program to predict genes by comparing anonymous genomic sequences from two different species. It combines TBLASTX, a sequence similarity search program, with geneid, an *"ab initio"* gene prediction program. The latest version of SGP2 is downloadable from this site. A web server has been developed recently by Genís Parra, and it is available at http://genome.imim.es/software/sgp2/sgp2.html

http://genome.imim.es/software/sgp2/

SGP2 supplementary materials

Supplementary materials for the SGP2 paper [Parra *et al.*, 2003] are available from this section. SGP2 is a gene prediction program that combines *"ab initio"* gene prediction with TBLASTX searches between two genome sequences to provide both sensitive and specific gene predictions.

http://genome.imim.es/datasets/sgp2002/

UCSC GENOME BROWSER

This site contains the reference sequence and working draft assemblies for a large collection of genomes. It also shows the *CFTR* (cystic fibrosis) region in 13 species and provides a portal to the ENCODE project. The UCSC GENOME BROWSER zooms and scrolls over chromosomes, showing the work of annotators worldwide.

http://genome.ucsc.edu/

Bibliography

J.F. Abril, R. Castelo, and R. Guigó. Comparison of splice sites in mammals and chicken. *Genome Res*, 15(1):111–119, Jan 3 2005. *Published online before print in Dec 8, 2004.*

J.F. Abril and R. Guigó. `gff2ps`: visualizing genomic annotations. *Bioinformatics*, 16(8):743–4, Aug 2000.

J.F. Abril, R. Guigó, and T. Wiehe. `gff2aplot`: Plotting sequence comparisons. *Bioinformatics*, 19 (18):2477–2479, Dec 12 2003.

M.D. Adams, S.E. Celniker, R.A. Holt, C.A. Evans, J.D. Gocayne, P.G. Amanatides, S.E. Scherer, P.W. Li, R.A. Hoskins, R.F. Galle, R.A. George, S.E. Lewis, S. Richards, M. Ashburner, S.N. Henderson, and others (including **J.F. Abril**). The genome sequence of *Drosophila melanogaster*. *Science*, 287 (5461):2185–95, Mar 24 2000.

Adobe Systems Inc. *PostScript Language Reference Manual.* Addison-Wesley Publishing Company, Inc., third edition, March 1999. ISBN 0-201-37922-8.

M. Aebi, H. Hornig, and C. Weissmann. 5′ cleavage site in eukaryotic pre-mRNA splicing is determined by the overall 5′ splice region, not by the conserved 5′ `GU`. *Cell*, 50(2):237–46, Jul 17 1987.

M. Alexandersson, S. Cawley, and L. Pachter. `SLAM`: cross-species gene finding and alignment with a generalized pair hidden Markov model. *Genome Res*, 13(3):496–502, Mar 2003.

S.F. Altschul, W. Gish, W. Miller, E.W. Myers, and D.J. Lipman. Basic local alignment search tool. *J Mol Biol*, 215(3):403–10, Oct 5 1990.

S.F. Altschul, T.L. Madden, A.A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D.J. Lipman. Gapped `BLAST` and `PSI-BLAST`: a new generation of protein database search programs. *Nucleic Acids Res*, 25(17):3389–402, Sep 1 1997.

F. Antequera and A. Bird. Number of CpG islands and genes in human and mouse. *Proc Natl Acad Sci U S A*, 90(24):11995–9, Dec 15 1993.

J.L. Ashurst, C.K. Chen, J.G. Gilbert, K. Jekosch, S. Keenan, P. Meidl, S.M. Searle, J. Stalker, R. Storey, S. Trevanion, L. Wilming, and T. Hubbard. The VErtebrate Genome Annotation (VEGA) database. *Nucleic Acids Res*, 33 Database Issue:D459–65, Jan 1 2005.

V.N. Babenko, P.S. Kosarev, O.V. Vishnevsky, V.G. Levitsky, V.V. Basin, and A.S. Frolov. Investigating extended regulatory regions of genomic DNA sequences. *Bioinformatics*, 15(7-8):644–53, Jul-Aug 1999.

V. Bafna and D.H. Huson. "The conserved exon method for gene finding.". In *Proc Int Conf Intell Syst Mol Biol*, volume 8, pages 3–12, 2000.

R.E. Banks, M.J. Dunn, D.F. Hochstrasser, J.C. Sanchez, W. Blackstock, D.J. Pappin, and P.J. Selby. Proteomics: new perspectives, new biomedical opportunities. *Lancet*, 356(9243):1749–56, Nov 18 2000.

E. Barillot, S. Pook, F. Guyon, C. Cussat-Blanc, E. Viara, and G. Vaysseix. The HUGEMAP Database: interconnection and visualization of human genome maps. *Nucleic Acids Res*, 27(1):119–22, Jan 1 1999.

S. Batzoglou, L. Pachter, J.P. Mesirov, B. Berger, and E.S. Lander. Human and mouse gene structure: comparative analysis and application to exon prediction. *Genome Res*, 10(7):950–8, Jul 2000.

E. Beitz. TEXshade: shading and labeling of multiple sequence alignments using LaTeX 2_ε. *Bioinformatics*, 16(2):135–9, Feb 2000.

S.M. Berget, C. Moore, and P.A. Sharp. Spliced segments at the 5′ terminus of adenovirus 2 late mRNA. *Proc Natl Acad Sci U S A*, 74(8):3171–5, Aug 1977.

E. Birney, D. Andrews, P. Bevan, M. Caccamo, G. Cameron, Y. Chen, L. Clarke, G. Coates, T. Cox, J. Cuff, V. Curwen, T. Cutts, T. Down, R. Durbin, E. Eyras, et al. ENSEMBL 2004. *Nucleic Acids Res*, 32(1):D468–70, Jan 1 2004a.

E. Birney, M. Clamp, and R. Durbin. GeneWise and Genomewise. *Genome Res*, 14(5):988–95, May 2004b.

E. Birney and R. Durbin. Dynamite: a flexible code generating language for dynamic programming methods used in sequence comparison. *Proc Int Conf Intell Syst Mol Biol*, 5:56–64, 1997.

D.L. Black. Mechanisms of alternative pre-messenger RNA splicing. *Annu Rev Biochem*, 72:291–336, 2003.

F.R. Blattner and J.L. Schroeder. A computer package for DNA sequence analysis. *Nucleic Acids Res*, 12(1 Pt 2):615–7, Jan 11 1984.

M. Blaxter, J. Daub, D. Guiliano, J. Parkinson, and C. Whitton. The *Brugia malayi* genome project: expressed sequence tags and gene discovery. *Trans R Soc Trop Med Hyg*, 96(1):7–17, Jan-Feb 2002.

P. Blayo, P. Rouzé, and M.-F. Sagot. Orphan gene finding - An exon assembly approach. *Theoretical Computer Science*, 290(3):1407–1431, 2002.

D. Boffelli, J. McAuliffe, D. Ovcharenko, K.D. Lewis, I. Ovcharenko, L. Pachter, and E.M. Rubin. Phylogenetic shadowing of primate sequences to find functional regions of the human genome. *Science*, 299(5611):1391–4, Feb 28 2003.

M. Borodovsky and J. McIninch. GeneMark: Parallel gene recognition for both DNA strands. *Computer and Chemistry*, 17:123–134, 1993.

M.R. Brent and R. Guigó. Recent advances in gene structure prediction. *Curr Opin Struct Biol*, 14(3): 264–72, Jun 2004.

D. Brett, H. Pospisil, J. Valcarcel, J. Reich, and P. Bork. Alternative splicing and genome complexity. *Nat Genet*, 30(1):29–30, Jan 2002.

C.T. Brown, A.G. Rust, P.J. Clarke, Z. Pan, M.J. Schilstra, T. De Buysscher, G. Griffin, B.J. Wold, R.A. Cameron, E.H. Davidson, and H. Bolouri. New computational approaches for analysis of cis-regulatory networks. *Dev Biol*, 246(1):86–102, Jun 1 2002.

A.R. Buchman and P. Berg. Comparison of intron-dependent and intron-independent gene expression. *Mol Cell Biol*, 8(10):4395–405, Oct 1988.

C.J. Bult, J.A. Blake, J.E. Richardson, J.A. Kadin, J.T. Eppig, R.M. Baldarelli, K. Barsanti, M. Baya, J.S. Beal, W.J. Boddy, D.W. Bradt, D.L. Burkart, N.E. Butler, J. Campbell, R. Corey, et al. The Mouse Genome Database (MGD): integrating biology with the genome. *Nucleic Acids Res*, 32 Database issue: D476–81, Jan 1 2004.

C. Burge and S. Karlin. Prediction of complete gene structures in human genomic DNA. *J Mol Biol*, 268(1):78–94, Apr 25 1997.

C. B. Burge, T. Tuschl, and P. S. Sharp. *The RNA world*, volume 37 of *Cold Spring Harbor Monograph Series*, chapter "Splicing Precursors to mRNAs by the Spliceosomes.", pages 525–560. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York, USA, 2nd edition, 1999. ISBN 0-87969-589-7.

C.B. Burge, R.A. Padgett, and P.A. Sharp. Evolutionary fates and origins of U12-type introns. *Mol Cell*, 2(6):773–85, Dec 1998.

M. Burset and R. Guigó. Evaluation of gene structure prediction programs. *Genomics*, 34(3):353–67, Jun 15 1996.

M. Burset, I.A. Seledtsov, and V.V. Solovyev. Analysis of canonical and non-canonical splice sites in mammalian genomes. *Nucleic Acids Res*, 28(21):4364–75, Nov 1 2000.

S.B. Cannon, A. Kozik, B. Chan, R. Michelmore, and N.D. Young. DiagHunter and GenoPix2D: programs for genomic comparisons, large-scale homology discovery and visualization. *Genome Biol*, 4(10):R68, 2003.

L. Cartegni, S.L. Chew, and A.R. Krainer. Listening to silence and understanding nonsense: exonic mutations that affect splicing. *Nat Rev Genet*, 3(4):285–98, Apr 2002.

R. Castelo, G. Parra, and R. Guigó. exstral: EXon STRucture over an ALignment. *unpublished* 2004.

M. Chagoyen, M.E. Kurul, P.A. De-Alarcon, J.M. Carazo, and A. Gupta. Designing and executing scientific workflows with a programmable integrator. *Bioinformatics*, 20(13):2092–100, Sep 1 2004.

K. Chakrabarti and L. Pachter. Visualization of multiple genome annotations and alignments with the K-BROWSER. *Genome Res*, 14(4):716–20, Apr 2004.

J. Cheung, X. Estivill, R. Khaja, J.R. MacDonald, K. Lau, L.C. Tsui, and S.W. Scherer. Genome-wide detection of segmental duplications and potential assembly errors in the human genome sequence. *Genome Biol*, 4(4):R25, 2003.

F. Chiaromonte, S. Yang, L. Elnitski, V.B. Yap, W. Miller, and R.C. Hardison. Association between divergence and interspersed repeats in mammalian noncoding genomic DNA. *Proc Natl Acad Sci U S A*, 98(25):14503–8, Dec 4 2001.

L.T. Chow, R.E. Gelinas, T.R. Broker, and R.J. Roberts. An amazing sequence arrangement at the 5′ ends of adenovirus 2 messenger RNA. *Cell*, 12(1):1–8, Sep 1977.

K.R. Christie, S. Weng, R. Balakrishnan, M.C. Costanzo, K. Dolinski, S.S. Dwight, S.R. Engel, B. Feierbach, D.G. Fisk, J.E. Hirschman, E.L. Hong, L. Issel-Tarver, R. Nash, A. Sethuraman, B. Starr, et al. *Saccharomyces* Genome Database (SGD) provides tools to identify and analyze sequences from *Saccharomyces cerevisiae* and related sequences from other organisms. *Nucleic Acids Res*, 32(1):D311–4, Jan 1 2004.

T.J. Chuang, W.C. Lin, H.C. Lee, C.W. Wang, K.L. Hsiao, Z.H. Wang, D. Shieh, S.C. Lin, and L.Y. Ch'ang. A complexity reduction algorithm for analysis and annotation of large genomic sequences. *Genome Res*, 13(2):313–22, Feb 2003.

M. Clamp, D. Andrews, D. Barker, P. Bevan, G. Cameron, Y. Chen, L. Clark, T. Cox, J. Cuff, V. Curwen, T. Down, R. Durbin, E. Eyras, J. Gilbert, M. Hammond, et al. ENSEMBL 2002: accommodating comparative genomics. *Nucleic Acids Res*, 31(1):38–42, Jan 1 2003.

J.E. Cleaver, C. Collins, J. Ellis, and S. Volik. Genome sequence and splice site analysis of low-fidelity DNA polymerases H and I involved in replication of damaged DNA. *Genomics*, 82(5):561–70, Nov 2003.

C.A. Collins and C. Guthrie. The question remains: is the spliceosome a ribozyme? *Nat Struct Biol*, 7 (10):850–4, Oct 2000.

F.S. Collins, E.D. Green, A.E. Guttmacher, and M.S. Guyer. A vision for the future of genomics research. *Nature*, 422(6934):835–47, Apr 24 2003.

J.W. Conaway, A. Shilatifard, A. Dvir, and R.C. Conaway. Control of elongation by RNA polymerase II. *Trends Biochem Sci*, 25(8):375–80, Aug 2000.

J. Corden and C. Ingles. *Transccriptional Regulation*, chapter "Carboxy-terminal domain of the largest subunit of eukaryotic RNA polymerase II", pages 81–108. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY (USA), 1992.

A. Cornish-Bowden. Nomenclature for incompletely specified bases in nucleic acid sequences: recommendations 1984. *Nucleic Acids Res*, 13(9):3021–30, May 10 1985.

E. Coward, S.A. Haas, and M. Vingron. SpliceNest: visualizing gene structure and alternative splicing based on EST clusters. *Trends Genetics*, 18(1):53–55, 2002.

P. Cramer, C.G. Pesce, F.E. Baralle, and A.R. Kornblihtt. Functional association between promoter structure and transcript alternative splicing. *Proc Natl Acad Sci U S A*, 94(21):11456–60, Oct 14 1997.

V. Curwen, E. Eyras, T.D. Andrews, L. Clarke, E. Mongin, S.M. Searle, and M. Clamp. The ENSEMBL automatic gene annotation system. *Genome Res*, 14(5):942–50, May 2004.

B. Datta and A.M. Weiner. Genetic evidence for base pairing between U2 and U6 snRNA in mammalian mRNA splicing. *Nature*, 352(6338):821–4, Aug 29 1991.

M. de la Mata, C.R. Alonso, S. Kadener, J.P. Fededa, M. Blaustein, F. Pelisch, P. Cramer, D. Bentley, and A.R. Kornblihtt. A slow RNA polymerase II affects alternative splicing *in vivo*. *Mol Cell*, 12(2): 525–32, Aug 2003.

A.L. Delcher, S. Kasif, R.D. Fleischmann, J. Peterson, O. White, and S.L. Salzberg. Alignment of whole genomes. *Nucleic Acids Res*, 27(11):2369–76, Jun 1 1999.

E.T. Dermitzakis, A. Reymond, R. Lyle, N. Scamuffa, C. Ucla, S. Deutsch, B.J. Stevenson, V. Flegel, P. Bucher, C.V. Jongeneel, and S.E. Antonarakis. Numerous potentially functional but non-genic conserved sequences on human chromosome 21. *Nature*, 420(6915):578–82, Dec 5 2002.

J. Devereux, P. Haeberli, and O. Smithies. A comprehensive set of sequence analysis programs for the VAX. *Nucleic Acids Res*, 12(1 Pt 1):387–95, Jan 11 1984.

C. Dewey, J.Q. Wu, S. Cawley, M. Alexandersson, R. Gibbs, and L. Pachter. Accurate identification of novel human genes through simultaneous gene prediction in human, mouse, and rat. *Genome Res*, 14(4):661–4, Apr 2004.

R.C. Dietrich, R. Incorvaia, and R.A. Padgett. Terminal intron dinucleotide sequences do not distinguish between U2- and U12-dependent introns. *Mol Cell*, 1(1):151–60, Dec 1997.

S. Dong and D.B. Searls. Gene structure prediction by linguistic methods. *Genomics*, 23(3):540–51, Oct 1994.

R.D. Dowell, R.M. Jokerst, A. Day, S.R. Eddy, and L. Stein. The distributed annotation system. *BMC Bioinformatics*, 2(1):7, 2001.

I. Dubchak, M. Brudno, G.G. Loots, L. Pachter, C. Mayor, E.M. Rubin, and K.A. Frazer. Active conservation of noncoding sequences revealed by three-way species comparisons. *Genome Res*, 10(9): 1304–6, Sep 2000.

I. Dunham, N. Shimizu, B.A. Roe, S. Chissoe, A.R. Hunt, J.E. Collins, R. Bruskiewich, D.M. Beare, M. Clamp, L.J. Smink, R. Ainscough, J.P. Almeida, A. Babbage, C. Bagguley, J. Bailey, et al. The DNA sequence of human chromosome 22. *Nature*, 402(6761):489–95, Dec 2 1999.

R. Durbin, S. Eddy, A. Crogh, and G. Mitchison. *Biological Sequence Analysis: Probabilistic Models of Protein and Nucleic Acids*. Cambridge University Press, first edition, 1998. ISBN 0-521-62971-3.

R. Durbin and J. Thierry-Mieg. The ACEDB genome database. URL `http://www.acedb.org/`. *unpublished* 1993.

L. Duret, E. Gasteiger, and G. Perriere. `LALNVIEW`: a graphical viewer for pairwise sequence alignments. *Comput Appl Biosci*, 12(6):507–10, Dec 1996.

I. Ebersberger, D. Metzler, C. Schwarz, and S. Paabo. Genomewide comparison of DNA sequences between humans and chimpanzees. *Am J Hum Genet*, 70(6):1490–7, Jun 2002.

J.H. Edwards. The `Oxford Grid`. *Ann Hum Genet*, 55 (Pt 1):17–31, Jan 1991.

Y.J. Edwards, T.J. Carver, T. Vavouri, M. Frith, M.J. Bishop, and G. Elgar. `Theatre`: A software tool for detailed comparative analysis and visualization of genomic sequence. *Nucleic Acids Res*, 31(13): 3510–7, Jul 1 2003.

F.H. Eeckman and R. Durbin. `ACeDB` and `macace`. *Methods Cell Biol*, 48:583–605, 1995.

ENCODE Project Consortium. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science*, 306 (5696):636–40, Oct 22 2004.

X. Estivill, J. Cheung, M.A. Pujana, K. Nakabayashi, S.W. Scherer, and L.C. Tsui. Chromosomal regions containing high-density and ambiguously mapped putative single nucleotide polymorphisms (SNPs) correlate with segmental duplications in the human genome. *Hum Mol Genet*, 11 (17):1987–95, Aug 15 2002.

T. Etzold and P. Argos. `SRS`—an indexing and retrieval tool for flat file data libraries. *Comput Appl Biosci*, 9(1):49–57, Feb 1993.

B. Ewing and P. Green. Analysis of expressed sequence tags indicates 35,000 human genes. *Nat Genet*, 25(2):232–4, Jun 2000.

Z. Fang, M. Polacco, S. Chen, S. Schroeder, D. Hancock, H. Sanchez, and E. Coe. `cMap`: the comparative genetic map viewer. *Bioinformatics*, 19(3):416–7, Feb 12 2003.

A. Fedorov, A.F. Merican, and W. Gilbert. Large-scale comparison of intron positions among animal, plant, and fungal genes. *Proc Natl Acad Sci U S A*, 99(25):16128–33, Dec 10 2002.

E.S. Ferlanti, J.F. Ryan, I. Makalowska, and A.D. Baxevanis. `WebBLAST` 2.0: an integrated solution for organizing and analyzing sequence data. *Bioinformatics*, 15(5):422–3, May 1999.

J.W. Fickett. Recognition of protein coding regions in DNA sequences. *Nucleic Acids Res*, 10(17): 5303–18, Sep 11 1982.

C. Fields, M.D. Adams, O. White, and J.C. Venter. How many genes in the human genome? *Nat Genet*, 7(3):345–6, Jul 1994.

C.A. Fields and C.A. Soderlund. `gm`: a practical tool for automating DNA sequence analysis. *Comput Appl Biosci*, 6(3):263–70, Jul 1990.

S. Fischer, J. Crabtree, B. Brunk, M. Gibson, and G.C. Overton. `bioWidgets`: data interaction components for genomics. *Bioinformatics*, 15(10):837–46, Oct 1999.

W.M. Fitch. An improved method of testing for evolutionary homology. *J Mol Biol*, 16(1):9–16, Mar 1966.

P. Flicek, E. Keibler, P. Hu, I. Korf, and M.R. Brent. Leveraging the mouse genome for gene prediction in human: from whole-genome shotgun reads to a global synteny map. *Genome Res*, 13(1):46–54, Jan 2003.

L. Florea, G. Hartzell, Z. Zhang, G.M. Rubin, and W. Miller. A computer program for aligning a cDNA sequence with a genomic DNA sequence. *Genome Res*, 8(9):967–74, Sep 1998.

L. Florea, M. McClelland, C. Riemer, S. Schwartz, and W. Miller. `EnteriX` 2003: Visualization tools for genome alignments of Enterobacteriaceae. *Nucleic Acids Res*, 31(13):3527–32, Jul 1 2003.

A. Fortna and K. Gardiner. Genomic sequence analysis tools: a user's guide. *Trends Genet*, 17(3): 158–64, Mar 2001.

K.A. Frazer, L. Elnitski, D.M. Church, I. Dubchak, and R.C. Hardison. Cross-species sequence comparisons: a review of methods and available resources. *Genome Res*, 13(1):1–12, Jan 2003.

X.D. Fu. Towards a splicing code. *Cell*, 119(6):736–8, Dec 17 2004.

M.S. Gelfand. Computer prediction of the exon-intron structure of mammalian pre-mRNAs. *Nucleic Acids Res*, 18(19):5865–9, Oct 11 1990.

M.S. Gelfand, A.A. Mironov, and P.A. Pevzner. Gene recognition via spliced sequence alignment. *Proc Natl Acad Sci U S A*, 93(17):9061–6, Aug 20 1996.

A.J. Gibbs and G.A. McIntyre. The `diagram`, a method for comparing sequences. Its use with amino acid and nucleotide sequences. *Eur J Biochem*, 16(1):1–11, Sep 1970.

R.A. Gibbs, G.M. Weinstock, M.L. Metzker, D.M. Muzny, E.J. Sodergren, S. Scherer, G. Scott, D. Steffen, K.C. Worley, P.E. Burch, G. Okwuonu, S. Hines, L. Lewis, C. DeRamo, O. Delgado, and others (Rat Genome Sequencing Project Consortium, RGSPC; including **J.F. Abril**). Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature*, 428(6982):493–521, Apr 1 2004.

R. Gibson and D.R. Smith. Genome visualization made fast and simple. *Bioinformatics*, 19(11):1449–50, Jul 22 2003.

R. Gil, F.J. Silva, E. Zientz, F. Delmotte, F. Gonzalez-Candelas, A. Latorre, C. Rausell, J. Kamerbeek, J. Gadau, B. Holldobler, R.C. van Ham, R. Gross, and A. Moya. The genome sequence of *Blochmannia floridanus*: comparative analysis of reduced genomes. *Proc Natl Acad Sci U S A*, 100(16):9388–93, Aug 5 2003.

D.G. Gilbert. `euGenes`: a eukaryote genome information system. *Nucleic Acids Res*, 30(1):145–8, Jan 1 2002.

P. Gilligan, S. Brenner, and B. Venkatesh. *Fugu* and human sequence comparison identifies novel human genes and conserved non-coding sequences. *Gene*, 294(1-2):35–44, Jul 10 2002.

W. Gish. Washington University BLAST. URL http://blast.wustl.edu. *unpublished* 1996–2004.

J.D. Glasner, P. Liss, 3.r.d. Plunkett G, A. Darling, T. Prasad, M. Rusch, A. Byrnes, M. Gilson, B. Biehl, F.R. Blattner, and N.T. Perna. ASAP, a systematic annotation package for community analysis of genomes. *Nucleic Acids Res*, 31(1):147–51, Jan 1 2003.

A.C. Goldstrohm, A.L. Greenleaf, and M.A. Garcia-Blanco. Co-transcriptional splicing of pre-messenger RNAs: considerations for the mechanism of alternative splicing. *Gene*, 277(1-2):31–47, Oct 17 2001.

N. Goodman. Biological data becomes computer literate: new advances in bioinformatics. *Curr Opin Biotechnol*, 13(1):68–71, Feb 2002.

B. Göttgens, L.M. Barton, J.G. Gilbert, A.J. Bench, M.J. Sanchez, S. Bahn, S. Mistry, D. Grafham, A. McMurray, M. Vaudin, E. Amaya, D.R. Bentley, A.R. Green, and A.M. Sinclair. Analysis of vertebrate *SCL* loci identifies conserved enhancers. *Nat Biotechnol*, 18(2):181–6, Feb 2000.

B. Göttgens, J.G. Gilbert, L.M. Barton, D. Grafham, J. Rogers, D.R. Bentley, and A.R. Green. Long-range comparison of human and mouse *SCL* loci: localized regions of sensitivity to restriction endonucleases correspond precisely with peaks of conserved noncoding sequences. *Genome Res*, 11(1):87–97, Jan 2001.

E. Graziano and P. Arus. FITMAPS and SHOWMAP: two programs for graphical comparison and plotting of genetic maps. *J Hered*, 93(3):225–7, May-Jun 2002.

A.L. Greenleaf. Positive patches and negative noodles: linking RNA processing to transcription? *Trends Biochem Sci*, 18(4):117–9, Apr 1993.

R. Guigo. Assembling genes from predicted exons in linear time with dynamic programming. *J Comput Biol*, 5(4):681–702, Winter 1998.

R. Guigó. *Genetic Databases.*, chapter "DNA Composition, Codon Usage and Exon Prediction.", pages 53–80. Academic Press, San Diego, California, USA, 1999. ISBN 0-12-101625-0.

R. Guigó, P. Agarwal, **J.F. Abril**, M. Burset, and J.W. Fickett. An assessment of gene prediction accuracy in large DNA sequences. *Genome Res*, 10(10):1631–42, Oct 2000.

R. Guigó, E.T. Dermitzakis, P. Agarwal, C.P. Ponting, G. Parra, A. Reymond, **J.F. Abril**, E. Keibler, R. Lyle, C. Ucla, S.E. Antonarakis, and M.R. Brent. Comparison of mouse and human genomes followed by experimental verification yields an estimated 1,019 additional genes. *Proc Natl Acad Sci U S A*, 100(3):1140–5, Feb 4 2003.

R. Guigó, S. Knudsen, N. Drake, and T. Smith. Prediction of gene structure. *J Mol Biol*, 226(1):141–57, Jul 5 1992.

R. Guigó and M.Q. Zhang. *Mammalian Genomics.*, chapter "Gene predictions and Annotations.", page (*in press*). CAB International, 2004. ISBN 0-851-99910-7.

C. Gybas and P. Jambeck. *Developing Bioinformatics Computer Skills.* O'Reilly & Associates, Inc., first edition, April 2003. ISBN 1-56592-664-1.

S.L. Hall and R.A. Padgett. Conserved sequences in a class of rare eukaryotic nuclear introns with non-consensus splice sites. *J Mol Biol*, 239(3):357–65, Jun 10 1994.

S.L. Hall and R.A. Padgett. Requirement of U12 snRNA for in vivo splicing of a minor class of eukaryotic nuclear pre-mRNA introns. *Science*, 271(5256):1716–8, Mar 22 1996.

R.C. Hardison. Conserved noncoding sequences are reliable guides to regulatory elements. *Trends Genet*, 16(9):369–72, Sep 2000.

R.C. Hardison, J. Oeltjen, and W. Miller. Long human-mouse sequence alignments reveal novel regulatory elements: a reason to sequence the mouse genome. *Genome Res*, 7(10):959–66, Oct 1997.

M.P. Hare and S.R. Palumbi. High intron sequence conservation across three mammalian orders suggests functional constraints. *Mol Biol Evol*, 20(6):969–78, Jun 2003.

T.W. Harris, N. Chen, F. Cunningham, M. Tello-Ruiz, I. Antoshechkin, C. Bastiani, T. Bieri, D. Blasiar, K. Bradnam, J. Chan, C.K. Chen, W.J. Chen, P. Davis, E. Kenny, R. Kishore, et al. WormBase: a multi-species resource for nematode biology and genomics. *Nucleic Acids Res*, 32(1):D411–7, Jan 1 2004.

P.M. Harrison, A. Kumar, N. Lang, M. Snyder, and M. Gerstein. A question of size: the eukaryotic proteome and the problems in defining it. *Nucleic Acids Res*, 30(5):1083–90, Mar 1 2002.

M.L. Hastings and A.R. Krainer. Pre-mRNA splicing in the new millennium. *Curr Opin Cell Biol*, 13 (3):302–9, Jun 2001.

D.L. Hatfield and V.N. Gladyshev. How selenium has altered our understanding of the genetic code. *Mol Cell Biol*, 22(11):3565–76, Jun 2002.

M. Hattori, A. Fujiyama, T.D. Taylor, H. Watanabe, T. Yada, H.S. Park, A. Toyoda, K. Ishii, Y. Totoki, D.K. Choi, Y. Groner, E. Soeda, M. Ohki, T. Takagi, Y. Sakaki, et al. The DNA sequence of human chromosome 21. *Nature*, 405(6784):311–9, May 18 2000.

T.P. Hausner, L.M. Giglio, and A.M. Weiner. Evidence for base-pairing between mammalian U2 and U6 small nuclear ribonucleoprotein particles. *Genes Dev*, 4(12A):2146–56, Dec 1990.

J. Healy, E.E. Thomas, J.T. Schwartz, and M. Wigler. Annotating large genomes with exact word matches. *Genome Res*, 13(10):2306–15, Oct 2003.

S. Heber, M. Alekseyev, S.H. Sze, H. Tang, and P.A. Pevzner. Splicing graphs and EST assembly problem. *Bioinformatics*, 18 Suppl 1:S181–8, Jul 2002.

J. Henderson, S. Salzberg, and K.H. Fasman. Finding genes in DNA with a Hidden Markov Model. *J Comput Biol*, 4(2):127–41, Summer 1997.

M.W. Hentze and A.E. Kulozik. A perfect message: RNA surveillance and nonsense-mediated decay. *Cell*, 96(3):307–10, Feb 5 1999.

C. Hertz-Fowler, C.S. Peacock, V. Wood, M. Aslett, A. Kerhornou, P. Mooney, A. Tivey, M. Berriman, N. Hall, K. Rutherford, J. Parkhill, A.C. Ivens, M.A. Rajandream, and B. Barrell. GeneDB: a resource for prokaryotic and eukaryotic organisms. *Nucleic Acids Res*, 32 Database issue:D339–43, Jan 1 2004.

L.W. Hillier, W. Miller, E. Birney, W. Warren, R.C. Hardison, C.P. Ponting, P. Bork, D.W. Burt, M.A. Groenen, M.E. Delany, J.B. Dodgson, G. Fingerprint Map Sequence, Assembly, A.T. Chinwalla, P.F. Cliften, S.W. Clifton, and others (International Chicken Genome Sequencing Consortium, ICGSC; including **J.F. Abril**). Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature*, 432(7018):695–716, Dec 9 2004.

H. Le Hir, E. Izaurralde, L.E. Maquat, and M.J. Moore. The spliceosome deposits multiple proteins 20-24 nucleotides upstream of mRNA exon-exon junctions. *EMBO J*, 19(24):6860–9, Dec 15 2000.

J.B. Hogenesch, K.A. Ching, S. Batalov, A.I. Su, J.R. Walker, Y. Zhou, S.A. Kay, P.G. Schultz, and M.P. Cooke. A comparison of the CELERA and ENSEMBL predicted gene sets reveals little overlap in novel genes. *Cell*, 106(4):413–5, Aug 24 2001.

R.A. Holt, G.M. Subramanian, A. Halpern, G.G. Sutton, R. Charlab, D.R. Nusskern, P. Wincker, A.G. Clark, J.M. Ribeiro, R. Wides, S.L. Salzberg, B. Loftus, M. Yandell, W.H. Majoros, D.B. Rusch, and others (including **J.F. Abril**). The genome sequence of the malaria mosquito *Anopheles gambiae*. *Science*, 298(5591):129–49, Oct 4 2002.

S. Hoon, K.K. Ratnapu, J.M. Chia, B. Kumarasamy, X. Juguang, M. Clamp, A. Stabenau, S. Potter, L. Clarke, and E. Stupka. Biopipe: a flexible framework for protocol-based bioinformatics analysis. *Genome Res*, 13(8):1904–15, Aug 2003.

K.J. Howe, C.M. Kane, and J.r. Ares M. Perturbation of transcription elongation influences the fidelity of internal exon inclusion in *Saccharomyces cerevisiae*. *RNA*, 9(8):993–1006, Aug 2003.

X. Huang and W. Miller. A time-efficient, linear-space local similarity algorithm. *Adv. Appl. Math.*, 12:337–357, 1991.

A.K. Hudek, J. Cheung, A.P. Boright, and S.W. Scherer. Genescript: DNA sequence annotation pipeline. *Bioinformatics*, 19(9):1177–8, Jun 12 2003.

G.B. Hutchinson and M.R. Hayden. The prediction of exons through an analysis of spliceable open reading frames. *Nucleic Acids Res*, 20(13):3453–62, Jul 11 1992.

R. Ierusalimschy, L. H. de Figueiredo, and W. Celes Filho. Lua—an extensible extension language. *Softw. Pract. Exper.*, 26(6):635–652, 1996.

R. Incorvaia and R.A. Padgett. Base pairing with U6atac snRNA is required for 5′ splice site activation of U12-dependent introns *in vivo*. *RNA*, 4(6):709–18, Jun 1998.

International Human Genome Sequencing Consortium, IHGSC. Finishing the euchromatic sequence of the human genome. *Nature*, 431(7011):931–45, Oct 21 2004.

Y. Ishigaki, X. Li, G. Serin, and L.E. Maquat. Evidence for a pioneer round of mRNA translation: mRNAs subject to nonsense-mediated decay in mammalian cells are bound by *CBP80* and *CBP20*. *Cell*, 106(5):607–17, Sep 7 2001.

IUPAC-IUB JCBN. IUPAC-IUB Joint Commission on Biochemical Nomenclature (JCBN). Nomenclature and symbolism for amino acids and peptides. Recommendations 1983. *Biochem J*, 219(2): 345–73, Apr 15 1984.

IUPAC-IUB JCBN. IUPAC-IUB Joint Commission on Biochemical Nomenclature (JCBN). Nomenclature and symbolism for amino acids and peptides. Corrections to recommendations 1983. *Eur J Biochem*, 213(1):2, Apr 1 1993.

I.J. Jackson. A reappraisal of non-consensus mRNA splice sites. *Nucleic Acids Res*, 19(14):3795–8, Jul 25 1991.

O. Jaillon, C. Dossat, R. Eckenberg, K. Eiglmeier, B. Segurens, J.M. Aury, C.W. Roth, C. Scarpelli, P.T. Brey, J. Weissenbach, and P. Wincker. Assessing the *Drosophila melanogaster* and *Anopheles gambiae* genome annotations using genome-wide sequence comparisons. *Genome Res*, 13(7):1595–9, Jul 2003.

D.C. Jamison. Open bioinformatics. *Bioinformatics*, 19(6):679–80, Apr 12 2003.

W. Jang, A. Hua, S.V. Spilson, W. Miller, B.A. Roe, and M.H. Meisler. Comparative sequence of human and mouse BAC clones from the *mnd2* region of chromosome 2p13. *Genome Res*, 9(1):53–61, Jan 1999.

N. Jareborg and R. Durbin. Alfresco–a workbench for comparative genomic sequence analysis. *Genome Res*, 10(8):1148–57, Aug 2000.

A.G. Jegga, S.P. Sherwood, J.W. Carman, A.T. Pinski, J.L. Phillips, J.P. Pestian, and B.J. Aronow. Detection and visualization of compositionally similar cis-regulatory element clusters in orthologous and coordinately controlled genes. *Genome Res*, 12(9):1408–17, Sep 2002.

R.A. Jensen. Orthologs and paralogs - we need to get it right. *Genome Biol*, 2(8):INTERACTIONS1002, 2001.

K. Jungfer and P. Rodriguez-Tome. Mapplet: a CORBA-based genome map viewer. *Bioinformatics*, 14 (8):734–8, 1998.

D. Kampa, J. Cheng, P. Kapranov, M. Yamanaka, S. Brubaker, S. Cawley, J. Drenkow, A. Piccolboni, S. Bekiranov, G. Helt, H. Tammana, and T.R. Gingeras. Novel RNAs identified from an in-depth analysis of the transcriptome of human chromosomes 21 and 22. *Genome Res*, 14(3):331–42, Mar 2004.

D. Karolchik, R. Baertsch, M. Diekhans, T.S. Furey, A. Hinrichs, Y.T. Lu, K.M. Roskin, M. Schwartz, C.W. Sugnet, D.J. Thomas, R.J. Weber, D. Haussler, and W.J. Kent. The UCSC GENOME BROWSER Database. *Nucleic Acids Res*, 31(1):51–4, Jan 1 2003.

D. Karolchik, A.S. Hinrichs, T.S. Furey, K.M. Roskin, C.W. Sugnet, D. Haussler, and W.J. Kent. The UCSC TABLE BROWSER data retrieval tool. *Nucleic Acids Res*, 32(1):D493–6, Jan 1 2004.

L.P. Keegan, A. Gallo, and M.A. O'Connell. The many roles of an RNA editor. *Nat Rev Genet*, 2(11): 869–78, Nov 2001.

C. Keller, M. Corcoran, and R.J. Roberts. Computer programs for handling nucleic acid sequences. *Nucleic Acids Res*, 12(1 Pt 1):379–86, Jan 11 1984.

W.J. Kent. BLAT—the BLAST-like alignment tool. *Genome Res*, 12(4):656–64, Apr 2002.

W.J. Kent, C.W. Sugnet, T.S. Furey, K.M. Roskin, T.H. Pringle, A.M. Zahler, and D. Haussler. The human genome browser at UCSC. *Genome Res*, 12(6):996–1006, Jun 2002.

W.J. Kent and A.M. Zahler. Conservation, regulation, synteny, and introns in a large-scale *C. briggsae*-*C. elegans* genomic alignment. *Genome Res*, 10(8):1115–25, Aug 2000.

Paul Kitts. *The NCBI handbook [Internet]*, chapter Genome Assembly and Annotation Process. National Library of Medicine (US), National Center for Biotechnology Information, Bethesda (MD), October 2002. URL http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=handbook.chapter.1440.

J. Kling. Ultrafast DNA sequencing. *Nat Biotechnol*, 21(12):1425–7, Dec 2003.

I. Kolossova and R.A. Padgett. U11 snRNA interacts in vivo with the 5′ splice site of U12-dependent (AU-AC) pre-mRNA introns. *RNA*, 3(3):227–33, Mar 1997.

M.M. Konarska and P.A. Sharp. Interactions between small nuclear ribonucleoprotein particles in formation of spliceosomes. *Cell*, 49(6):763–74, Jun 19 1987.

I. Korf, P. Flicek, D. Duan, and M.R. Brent. Integrating genomic homology into gene structure prediction. *Bioinformatics*, 17 Suppl 1:S140–8, 2001.

A. Kozik, E. Kochetkova, and R. Michelmore. GenomePixelizer—a visualization program for comparative genomics within and between species. *Bioinformatics*, 18(2):335–6, Feb 2002.

A. Krause, S.A. Haas, E. Coward, and M. Vingron. SYSTERS, GeneNest, SpliceNest: exploring sequence space from genome to protein. *Nucleic Acids Res*, 30(1):299–300, Jan 1 2002.

A. Krogh. Two methods for improving performance of an HMM and their application for gene finding. *Proc Int Conf Intell Syst Mol Biol*, 5:179–86, 1997.

D. Kulp, D. Haussler, M. G. Reese, and F. H. Eeckman. "A Generalized Hidden Markov Model for the Recognition of Human Genes in DNA.". In D. J. States, P. Agarwal, T. Gaasterland, L. Hunter, and R. Smith, editors, *Proc Int Conf Intell Syst Mol Biol*, volume 4, pages 134–142, Menlo Park, California, 1996. AAAI press.

S. Kurtz, A. Phillippy, A.L. Delcher, M. Smoot, M. Shumway, C. Antonescu, and S.L. Salzberg. Versatile and open software for comparing large genomes. *Genome Biol*, 5(2):R12, 2004.

S. Kurtz and C. Schleiermacher. REPuter: fast computation of maximal repeats in complete genomes. *Bioinformatics*, 15(5):426–7, May 1999.

A.I. Lamond, M.M. Konarska, P.J. Grabowski, and P.A. Sharp. Spliceosome assembly involves the binding and release of U4 small nuclear ribonucleoprotein. *Proc Natl Acad Sci U S A*, 85(2):411–5, Jan 1988.

L. Lamport. *LaTeX A Document Preparation System.* Addison Wesley, second edition, 1994.

E.S. Lander, L.M. Linton, B. Birren, C. Nusbaum, M.C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, R. Funke, D. Gage, K. Harris, A. Heaford, J. Howland, and others (International Human Genome Sequencing Consortium, IHGSC). Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, Feb 15 2001.

A. Lefebvre, T. Lecroq, H. Dauchel, and J. Alexandre. FORRepeats: detects repeats on entire chromosomes and between genomes. *Bioinformatics*, 19(3):319–26, Feb 12 2003.

B. Lenhard, A. Sandelin, L. Mendoza, P. Engstrom, N. Jareborg, and W.W. Wasserman. Identification of conserved regulatory elements by comparative genome analysis. *J Biol*, 2(2):13, 2003.

C. Letondal. A Web interface generator for molecular biology programs in Unix. *Bioinformatics*, 17(1): 73–82, Jan 2001.

S.E. Lewis, S.M. Searle, N. Harris, M. Gibson, V. Lyer, J. Richter, C. Wiel, L. Bayraktaroglir, E. Birney, M.A. Crosby, J.S. Kaminker, B.B. Matthews, S.E. Prochnik, C.D. Smithy, J.L. Tupy, et al. Apollo: a sequence annotation editor. *Genome Biol*, 3(12):RESEARCH0082, 2002.

F. Liang, I. Holt, G. Pertea, S. Karamycheva, S.L. Salzberg, and J. Quackenbush. Gene index analysis of the human genome estimates approximately 120,000 genes. *Nat Genet*, 25(2):239–40, Jun 2000.

H.X. Liu, M. Zhang, and A.R. Krainer. Identification of functional exonic splicing enhancer motifs recognized by individual SR proteins. *Genes Dev*, 12(13):1998–2012, Jul 1 1998.

G.G. Loots, R.M. Locksley, C.M. Blankespoor, Z.E. Wang, W. Miller, E.M. Rubin, and K.A. Frazer. Identification of a coordinate regulator of *interleukins* 4, 13, and 5 by cross-species sequence comparisons. *Science*, 288(5463):136–40, Apr 7 2000.

S. Lu and B.R. Cullen. Analysis of the stimulatory effect of splicing on mRNA production and utilization in mammalian cells. *RNA*, 9(5):618–30, May 2003.

A.V. Lukashin and M. Borodovsky. GeneMark.hmm: new solutions for gene finding. *Nucleic Acids Res*, 26(4):1107–15, Feb 15 1998.

J. Lund, F. Chen, A. Hua, B. Roe, M. Budarf, B.S. Emanuel, and R.H. Reeves. Comparative sequence analysis of 634 kb of the mouse chromosome 16 region of conserved synteny with the human velocardiofacial syndrome region on chromosome 22q11.2. *Genomics*, 63(3):374–83, Feb 1 2000.

H.R. Luo, G.A. Moreau, N. Levin, and M.J. Moore. The human *Prp8* protein is a component of both U2- and U12-dependent spliceosomes. *RNA*, 5(7):893–908, Jul 1999.

H.D. Madhani and C. Guthrie. A novel base-pairing interaction between U2 and U6 snRNAs suggests a mechanism for the catalytic activation of the spliceosome. *Cell*, 71(5):803–17, Nov 27 1992.

E.M. Makarov, O.V. Makarova, H. Urlaub, M. Gentzel, C.L. Will, M. Wilm, and R. Luhrmann. Small nuclear ribonucleoprotein remodeling during catalytic activation of the spliceosome. *Science*, 298 (5601):2205–8, Dec 13 2002.

L. E. Maquat. *Translational Control of Gene Expression*, volume 39 of *Cold Spring Harbor Monograph Series*, chapter "Nonsense-mediated RNA decay in mammalian cells: a splicing-dependent means to down-regulate the levels of mRNAs that premature terminate translation.", pages 849–868. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY (USA), 2000. ISBN 0-87969-618-4.

L.E. Maquat. When cells stop making sense: effects of nonsense codons on RNA metabolism in vertebrate cells. *RNA*, 1(5):453–65, Jul 1995.

C. Mayor, M. Brudno, J.R. Schwartz, A. Poliakov, E.M. Rubin, K.A. Frazer, L.S. Pachter, and I. Dubchak. VISTA: visualizing global DNA sequence alignments of arbitrary length. *Bioinformatics*, 16(11):1046–7, Nov 2000.

T.S. McConnell, S.J. Cho, M.J. Frilander, and J.A. Steitz. Branchpoint selection in the splicing of U12-dependent introns *in vitro*. *RNA*, 8(5):579–86, May 2002.

S. McCracken, N. Fong, E. Rosonina, K. Yankulov, G. Brothers, D. Siderovski, A. Hessel, S. Foster, S. Shuman, and D.L. Bentley. 5′-Capping enzymes are targeted to pre-mRNA by binding to the phosphorylated carboxy-terminal domain of RNA polymerase II. *Genes Dev*, 11(24):3306–18, Dec 15 1997a.

S. McCracken, N. Fong, K. Yankulov, S. Ballantyne, G. Pan, J. Greenblatt, S.D. Patterson, M. Wickens, and D.L. Bentley. The C-terminal domain of RNA polymerase II couples mRNA processing to transcription. *Nature*, 385(6614):357–61, Jan 23 1997b.

I.M. Meyer and R. Durbin. Comparative *ab initio* prediction of gene structures using pair HMMs. *Bioinformatics*, 18(10):1309–18, Oct 2002.

I.M. Meyer and R. Durbin. Gene structure conservation aids similarity based gene prediction. *Nucleic Acids Res*, 32(2):776–83, 2004.

W. Miller. Comparison of genomic DNA sequences: solved and unsolved problems. *Bioinformatics*, 17(5):391–7, May 2001.

B. Modrek, A. Resch, C. Grasso, and C. Lee. Genome-wide detection of alternative splicing in expressed sequences of human genes. *Nucleic Acids Res*, 29(13):2850–9, Jul 1 2001.

S.B. Montgomery, T. Astakhova, M. Bilenky, E. Birney, T. Fu, M. Hassel, C. Melsopp, M. Rak, A.G. Robertson, M. Sleumer, A.S. Siddiqui, and S.J. Jones. Sockeye: a 3D environment for comparative genomics. *Genome Res*, 14(5):956–62, May 2004.

K.A. Montzka and J.A. Steitz. Additional low-abundance human small nuclear ribonucleoproteins: U11, U12, etc. *Proc Natl Acad Sci U S A*, 85(23):8885–9, Dec 1988.

R. Mott. EST_GENOME: a program to align spliced DNA sequences to unspliced genomic DNA. *Comput Appl Biosci*, 13(4):477–8, Aug 1997.

S.M. Mount. A catalogue of splice junction sequences. *Nucleic Acids Res*, 10(2):459–72, Jan 22 1982.

T. Mourier and D.C. Jeffares. Eukaryotic intron loss. *Science*, 300(5624):1393, May 30 2003.

C.J. Mungall, S. Misra, B.P. Berman, J. Carlson, E. Frise, N. Harris, B. Marshall, S. Shu, J.S. Kaminker, S.E. Prochnik, C.D. Smith, E. Smith, J.L. Tupy, C. Wiel, G.M. Rubin, et al. An integrated computational pipeline and database to support whole-genome sequence annotation. *Genome Biol*, 3(12): RESEARCH0081, 2002.

R.J. Mural, M.D. Adams, E.W. Myers, H.O. Smith, G.L. Miklos, R. Wides, A. Halpern, P.W. Li, G.G. Sutton, J. Nadeau, S.L. Salzberg, R.A. Holt, C.D. Kodira, F. Lu, L. Chen, et al. A comparison of whole-genome shotgun-derived mouse chromosome 16 and the human genome. *Science*, 296 (5573):1661–71, May 31 2002.

P. Nadkarni. Mapmerge: merge genomic maps. *Bioinformatics*, 14(4):310–6, 1998.

NCBI. Gnomon, predicting gene structures in genomic DNA. URL http://www.ncbi.nlm.nih.gov/genome/guide/gnomon.html. *unpublished* 2003.

A. Nekrutenko, W.Y. Chung, and W.H. Li. An evolutionary approach reveals a high protein-coding capacity of the human genome. *Trends Genet*, 19(6):306–10, Jun 2003.

A. Newman and C. Norman. Mutations in yeast U5 snRNA alter the specificity of 5′ splice-site cleavage. *Cell*, 65(1):115–23, Apr 5 1991.

A.J. Newman and C. Norman. U5 snRNA interacts with exon sequences at 5′ and 3′ splice sites. *Cell*, 68(4):743–54, Feb 21 1992.

A. Nott, S.H. Meislin, and M.J. Moore. A quantitative analysis of intron effects on mammalian gene expression. *RNA*, 9(5):607–17, May 2003.

J.C. Oeltjen, T.M. Malley, D.M. Muzny, W. Miller, R.A. Gibbs, and J.W. Belmont. Large-scale comparative sequence analysis of the human and murine Bruton's *tyrosine kinase* loci reveals conserved regulatory domains. *Genome Res*, 7(4):315–29, Apr 1997.

S.A. Olson. EMBOSS opens up sequence analysis. European Molecular Biology Open Software Suite. *Brief Bioinform*, 3(1):87–91, Mar 2002.

L.R. Otake, P. Scamborova, C. Hashimoto, and J.A. Steitz. The divergent U12-type spliceosome is required for pre-mRNA splicing and is essential for development in *Drosophila*. *Mol Cell*, 9(2): 439–46, Feb 2002.

I. Ovcharenko and G.G. Loots. Comparative genomic tools for exploring the human genome. *Cold Spring Harb Symp Quant Biol*, 68:283–91, 2003a.

I. Ovcharenko and G.G. Loots. Finding the Needle in the Haystack: Computational Strategies for Discovering Regulatory Sequences in Genomes. *Current Genomics*, 4(7):557–568, 2003b.

I. Ovcharenko, G.G. Loots, R.C. Hardison, W. Miller, and L. Stubbs. zPicture: dynamic alignment and visualization tool for analyzing conservation profiles. *Genome Res*, 14(3):472–7, Mar 2004a.

I. Ovcharenko, M.A. Nobrega, G.G. Loots, and L. Stubbs. `ECR Browser`: a tool for visualizing and accessing data from comparisons of multiple vertebrate genomes. *Nucleic Acids Res*, 32(Web Server issue):W280–6, Jul 1 2004b.

R. Overbeek, N. Larsen, T. Walunas, M. D'Souza, G. Pusch, J.r. Selkov E, K. Liolios, V. Joukov, D. Kaznadzey, I. Anderson, A. Bhattacharyya, H. Burd, W. Gardner, P. Hanke, V. Kapatral, et al. The `ERGO` genome analysis and discovery system. *Nucleic Acids Res*, 31(1):164–71, Jan 1 2003.

F. Pagani and F.E. Baralle. Genomic variants in exons and introns: identifying the splicing spoilers. *Nat Rev Genet*, 5(5):389–96, May 2004.

Q. Pan, M.A. Bakowski, Q. Morris, W. Zhang, B.J. Frey, T.R. Hughes, and B.J. Blencowe. Alternative splicing of conserved exons is frequently species-specific in human and mouse. *Trends Genet*, 21(2): 73–7, Feb 2005.

J. Parkinson and M. Blaxter. `SimiTri`—visualizing similarity relationships for groups of sequences. *Bioinformatics*, 19(3):390–5, Feb 12 2003.

G. Parra, P. Agarwal, **J.F. Abril**, T. Wiehe, J.W. Fickett, and R. Guigó. Comparative gene prediction in human and mouse. *Genome Res*, 13(1):108–17, Jan 2003.

J.D. Parsons. `Miropeats`: graphical DNA sequence comparisons. *Comput Appl Biosci*, 11(6):615–9, Dec 1995.

E. Passarge, B. Horsthemke, and R.A. Farber. Incorrect use of the term synteny. *Nat Genet*, 23(4):387, Dec 1999.

A.A. Patel and J.A. Steitz. Splicing double: insights from the second spliceosome. *Nat Rev Mol Cell Biol*, 4(12):960–70, Dec 2003.

W.R. Pearson and D.J. Lipman. Improved tools for biological sequence comparison. *Proc Natl Acad Sci U S A*, 85(8):2444–8, Apr 1988.

C.N.S. Pedersen and T. Scharl. "Comparative Methods for Gene Structure Prediction in Homologous Sequences.". In R. Guigó and D. Gusfield, editors, *"Algorithms in Bioinformatics": Proceedings of the Second International Workshop, WABI 2002*, volume 2452 of *Lecture Notes in Computer Science*, pages 220–234. Springer-Verlag, Berlin Heidelberg, 2002. ISBN 3-540-44211-1.

J.S. Pedersen and J. Hein. Gene finding with a hidden Markov model of genome structure and evolution. *Bioinformatics*, 19(2):219–27, Jan 22 2003.

L.A. Pennacchio. Insights from human/mouse genome comparisons. *Mamm Genome*, 14(7):429–36, Jul 2003.

L.A. Pennacchio, M. Olivier, J.A. Hubacek, J.C. Cohen, D.R. Cox, J.C. Fruchart, R.M. Krauss, and E.M. Rubin. An apolipoprotein influencing triglycerides in humans and mice revealed by comparative sequencing. *Science*, 294(5540):169–73, Oct 5 2001.

L.A. Pennacchio and E.M. Rubin. Genomic strategies to identify mammalian regulatory sequences. *Nat Rev Genet*, 2(2):100–9, Feb 2001.

L.A. Pennacchio and E.M. Rubin. Comparative genomic tools and databases: providing insights into the human genome. *J Clin Invest*, 111(8):1099–106, Apr 2003.

E. Pennisi. Bioinformatics. Gene counters struggle to get the right answer. *Science*, 301(5636):1040–1, Aug 22 2003.

S.C. Potter, L. Clarke, V. Curwen, S. Keenan, E. Mongin, S.M. Searle, A. Stabenau, R. Storey, and M. Clamp. The `Ensembl` analysis pipeline. *Genome Res*, 14(5):934–41, May 2004.

N.J. Proudfoot, A. Furger, and M.J. Dye. Integrating mRNA processing with transcription. *Cell*, 108 (4):501–12, Feb 22 2002.

K.D. Pruitt and D.R. Maglott. REFSEQ and LOCUSLINK: NCBI gene-centered resources. *Nucleic Acids Res*, 29(1):137–40, Jan 1 2001.

K.D. Pruitt, T. Tatusova, and D.R. Maglott. NCBI Reference Sequence (REFSEQ): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res*, 33 Database Issue:D501–4, Jan 1 2005.

J. Pustell and F.C. Kafatos. A convenient and adaptable package of DNA sequence analysis programs for microcomputers. *Nucleic Acids Res*, 10(1):51–9, Jan 11 1982.

W.C. Ray, J.r. Munson RS, and C.J. Daniels. `Tricross`: using dot-plots in sequence-id space to detect uncataloged intergenic features. *Bioinformatics*, 17(12):1105–12, Dec 2001.

R. Reed. Initial splice-site recognition and pairing during pre-mRNA splicing. *Curr Opin Genet Dev*, 6(2):215–20, Apr 1996.

M.G. Reese, G. Hartzell, N.L. Harris, U. Ohler, **J.F. Abril**, and S.E. Lewis. Genome annotation assessment in *Drosophila melanogaster*. *Genome Res*, 10(4):483–501, Apr 2000.

V.L. Reichert, H. Le Hir, M.S. Jurica, and M.J. Moore. 5′ exon interactions within the human spliceosome establish a framework for exon junction complex structure and assembly. *Genes Dev*, 16(21): 2778–91, Nov 1 2002.

K. Reichwald, J. Thiesen, T. Wiehe, J. Weitzel, W.A. Poustka, A. Rosenthal, M. Platzer, W.H. Stratling, and P. Kioschis. Comparative sequence analysis of the *MECP2*-locus in human and mouse reveals new transcribed regions. *Mamm Genome*, 11(3):182–90, Mar 2000.

Glenn C. Reid. *PostScript Language Program Design.* Addison-Wesley Publishing Company, Inc., twelfth edition, March 1996. ISBN 0-201-14396-8.

D. Reisman, E. Eaton, D. McMillin, N.A. Doudican, and K. Boggs. Cloning and characterization of murine *p53* upstream sequences reveals additional positive transcriptional regulatory elements. *Gene*, 274(1-2):129–37, Aug 22 2001.

S.Y. Rhee, W. Beavis, T.Z. Berardini, G. Chen, D. Dixon, A. Doyle, M. Garcia-Hernandez, E. Huala, G. Lander, M. Montoya, N. Miller, L.A. Mueller, S. Mundodi, L. Reiser, J. Tacklind, et al. The *Arabidopsis* Information Resource (`TAIR`): a model organism database providing a centralized, curated gateway to *Arabidopsis* biology, research materials and community. *Nucleic Acids Res*, 31(1):224–8, Jan 1 2003.

P. Rice, I. Longden, and A. Bleasby. `EMBOSS`: the European Molecular Biology Open Software Suite. *Trends Genet*, 16(6):276–7, Jun 2000.

H. Roest Crollius, O. Jaillon, A. Bernot, C. Dasilva, L. Bouneau, C. Fischer, C. Fizames, P. Wincker, P. Brottier, F. Quetier, W. Saurin, and J. Weissenbach. Estimate of human gene number provided by genome-wide analysis using *Tetraodon nigroviridis* DNA sequence. *Nat Genet*, 25(2):235–8, Jun 2000.

S. Rogic, A.K. Mackworth, and F.B. Ouellette. Evaluation of gene-finding programs on mammalian sequences. *Genome Res*, 11(5):817–32, May 2001.

I.B. Rogozin and Y.I. Pavlov. Theoretical analysis of mutation hotspots and their DNA sequence context specificity. *Mutat Res*, 544(1):65–85, Sep 2003.

S.W. Roy. Recent evidence for the exon theory of genes. *Genetica*, 118(2-3):251–66, Jul 2003.

K. Rutherford, J. Parkhill, J. Crook, T. Horsnell, P. Rice, M.A. Rajandream, and B. Barrell. `Artemis`: sequence visualization and annotation. *Bioinformatics*, 16(10):944–5, Oct 2000.

W.S. Ryu and J.E. Mertz. Simian virus 40 late transcripts lacking excisable intervening sequences are defective in both stability in the nucleus and transport to the cytoplasm. *J Virol*, 63(10):4386–94, Oct 1989.

A.A. Salamov and V.V. Solovyev. *Ab initio* gene finding in *Drosophila* genomic DNA. *Genome Res*, 10 (4):516–22, Apr 2000.

S. Salzberg, A.L. Delcher, K.H. Fasman, and J. Henderson. A decision tree system for finding genes in DNA. *J Comput Biol*, 5(4):667–80, Winter 1998.

A. Sandelin, W.W. Wasserman, and B. Lenhard. `ConSite`: web-based prediction of regulatory elements using cross-species comparison. *Nucleic Acids Res*, 32(Web Server issue):W249–52, Jul 1 2004.

N. Sato and S. Ehira. `GenoMap`, a circular genome data viewer. *Bioinformatics*, 19(12):1583–4, Aug 12 2003.

T.D. Schaal and T. Maniatis. Multiple distinct splicing enhancers in the protein-coding sequences of a constitutively spliced pre-mRNA. *Mol Cell Biol*, 19(1):261–73, Jan 1999.

C. Schneider, C.L. Will, O.V. Makarova, E.M. Makarov, and R. Luhrmann. Human U4/U6.U5 and U4atac/U6atac.U5 tri-snRNPs exhibit similar protein compositions. *Mol Cell Biol*, 22(10):3219–29, May 2002.

T.D. Schneider and R.M. Stephens. Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res*, 18(20):6097–100, Oct 25 1990.

S. Schwartz, L. Elnitski, M. Li, M. Weirauch, C. Riemer, A. Smit, E.D. Green, R.C. Hardison, and W. Miller. `MultiPipMaker` and supporting tools: Alignments and analysis of multiple genomic DNA sequences. *Nucleic Acids Res*, 31(13):3518–24, Jul 1 2003a.

S. Schwartz, W.J. Kent, A. Smit, Z. Zhang, R. Baertsch, R.C. Hardison, D. Haussler, and W. Miller. Human-mouse alignments with `BLASTZ`. *Genome Res*, 13(1):103–7, Jan 2003b.

S. Schwartz, W. Miller, C.M. Yang, and R.C. Hardison. Software tools for analyzing pairwise alignments of long sequences. *Nucleic Acids Res*, 19(17):4663–7, Sep 11 1991.

S. Schwartz, Z. Zhang, K.A. Frazer, A. Smit, C. Riemer, J. Bouck, R. Gibbs, R. Hardison, and W. Miller. `PipMaker`—a web server for aligning two genomic DNA sequences. *Genome Res*, 10(4):577–86, Apr 2000.

S.M. Searle, J. Gilbert, V. Iyer, and M. Clamp. The `otter` annotation system. *Genome Res*, 14(5): 963–70, May 2004.

D.B. Searls. Doing sequence analysis with your printer. *Comput Appl Biosci*, 9(4):421–6, Aug 1993.

D.B. Searls. `bioTk`: componentry for genome informatics graphical user interfaces. *Gene*, 163(2): GC1–16, Oct 3 1995.

P. Senapathy, M.B. Shapiro, and N.L. Harris. Splice junctions, branch point sites, and exons: sequence statistics, identification, and applications to genome project. *Methods Enzymol*, 183:252–78, 1990.

P.A. Sharp and C.B. Burge. Classification of introns: U2-type or U12-type. *Cell*, 91(7):875–9, Dec 26 1997.

J.C. Shepherd. Method to determine the reading frame of a protein from the purine/pyrimidine genome sequence and its possible evolutionary justification. *Proc Natl Acad Sci U S A*, 78(3):1596–600, Mar 1981.

S.T. Sherry, M.H. Ward, M. Kholodov, J. Baker, L. Phan, E.M. Smigielski, and K. Sirotkin. DBSNP: the NCBI database of genetic variation. *Nucleic Acids Res*, 29(1):308–11, Jan 1 2001.

A. Siepel and D. Haussler. Phylogenetic estimation of context-dependent substitution rates by maximum likelihood. *Mol Biol Evol*, 21(3):468–88, Mar 2004.

G.S. Slater and E. Birney. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics*, 6(1):31, Feb 15 2005.

A.F.A. Smit, R. Hubley, and P. Green. `RepeatMasker`. URL `http://www.repeatmasker.org/`. *unpublished* 1996–2004.

M.W. Smith. Structure of vertebrate genes: a statistical analysis implicating selection. *J Mol Evol*, 27 (1):45–55, 1988.

M.E. Smoot, S.A. Guerlain, and W.R. Pearson. Visualization of near optimal sequence alignments. *Bioinformatics*, Jan 29 2004.

E.E. Snyder and G.D. Stormo. Identification of coding regions in genomic DNA sequences: an application of dynamic programming and neural networks. *Nucleic Acids Res*, 21(3):607–13, Feb 11 1993.

V.V. Solovyev, A.A. Salamov, and C.B. Lawrence. Predicting internal exons by oligonucleotide composition and discriminant analysis of spliceable open reading frames. *Nucleic Acids Res*, 22(24): 5156–63, Dec 11 1994.

E.L. Sonnhammer and R. Durbin. A workbench for large-scale sequence homology analysis. *Comput Appl Biosci*, 10(3):301–7, Jun 1994.

E.L. Sonnhammer and R. Durbin. A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. *Gene*, 167(1-2):GC1–10, Dec 29 1995.

E.J. Sontheimer and J.A. Steitz. The U5 and U6 small nuclear RNAs as active site components of the spliceosome. *Science*, 262(5142):1989–96, Dec 24 1993.

R. Sorek and G. Ast. Intronic sequences flanking alternatively spliced exons are conserved between human and mouse. *Genome Res*, 13(7):1631–7, Jul 2003.

R. Staden. An interactive graphics program for comparing and aligning nucleic acid and amino acid sequences. *Nucleic Acids Res*, 10(9):2951–61, May 11 1982.

R. Staden. Computer methods to locate signals in nucleic acid sequences. *Nucleic Acids Res*, 12(1 Pt 2):505–19, Jan 11 1984a.

R. Staden. Graphic methods to determine the function of nucleic acid sequences. *Nucleic Acids Res*, 12(1 Pt 2):521–38, Jan 11 1984b.

R. Staden. The current status and portability of our sequence handling software. *Nucleic Acids Res*, 14(1):217–31, Jan 10 1986.

R. Staden. Methods to define and locate patterns of motifs in sequences. *Comput Appl Biosci*, 4(1): 53–60, Mar 1988.

R. Staden, K.F. Beal, and J.K. Bonfield. The Staden package, 1998. *Methods Mol Biol*, 132:115–30, 2000.

R. Staden and A.D. McLachlan. Codon preference and its use in identifying protein coding regions in long DNA sequences. *Nucleic Acids Res*, 10(1):141–56, Jan 11 1982.

J.E. Stajich, D. Block, K. Boulez, S.E. Brenner, S.A. Chervitz, C. Dagdigian, G. Fuellen, J.G. Gilbert, I. Korf, H. Lapp, H. Lehvaslaiho, C. Matsalla, C.J. Mungall, B.I. Osborne, M.R. Pocock, et al. The Bioperl toolkit: Perl modules for the life sciences. *Genome Res*, 12(10):1611–8, Oct 2002.

J. Stalker, B. Gibbins, P. Meidl, J. Smith, W. Spooner, H.R. Hotz, and A.V. Cox. The ENSEMBL Web site: mechanics of a genome browser. *Genome Res*, 14(5):951–5, May 2004.

M. Stanke and S. Waack. Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics*, 19 Suppl 2:II215–II225, Oct 2003.

L. Stein. Genome annotation: from sequence to biology. *Nat Rev Genet*, 2(7):493–503, Jul 2001.

L.D. Stein, S. Cartinhour, D. Thierry-Mieg, and J. Thierry-Mieg. JADE: an approach for interconnecting bioinformatics databases. *Gene*, 209(1-2):GC39–GC43, Mar 16 1998.

L.D. Stein, C. Mungall, S. Shu, M. Caudy, M. Mangone, A. Day, E. Nickerson, J.E. Stajich, T.W. Harris, A. Arva, and S. Lewis. The generic genome browser: a building block for a model organism system database. *Genome Res*, 12(10):1599–610, Oct 2002.

L.D. Stein and J. Thierry-Mieg. Scriptable access to the *Caenorhabditis elegans* genome sequence and other ACEDB databases. *Genome Res*, 8(12):1308–15, Dec 1998.

R. Stevens, C. Goble, P. Baker, and A. Brass. A classification of tasks in bioinformatics. *Bioinformatics*, 17(2):180–8, Feb 2001.

A. Stoltzfus, J.r. Logsdon JM, J.D. Palmer, and W.F. Doolittle. Intron "sliding" and the diversity of intron positions. *Proc Natl Acad Sci U S A*, 94(20):10739–44, Sep 30 1997.

G. Storz. An expanding universe of noncoding RNAs. *Science*, 296(5571):1260–3, May 17 2002.

C. Suter-Crazzolara and G. Kurapkat. An infrastructure for comparative genomics to functionally characterize genes and proteins. *Genome Inform Ser Workshop Genome Inform*, 11:24–32, 2000.

A.V. Sverdlov, V.N. Babenko, I.B. Rogozin, and E.V. Koonin. Preferential loss and gain of introns in $3'$ portions of genes suggests a reverse-transcription mechanism of intron insertion. *Gene*, 338(1): 85–91, Aug 18 2004.

A.V. Sverdlov, I.B. Rogozin, V.N. Babenko, and E.V. Koonin. Conservation versus parallel gains in intron evolution. *Nucleic Acids Res*, 33(6):1741–8, 2005.

A. Taneda. ADplot: detection and visualization of repetitive patterns in complete genomes. *Bioinformatics*, 20(5):701–8, Mar 22 2004.

W.Y. Tarn and J.A. Steitz. A novel spliceosome containing U11, U12, and U5 snRNPs excises a minor class (AT-AC) intron *in vitro*. *Cell*, 84(5):801–11, Mar 8 1996.

T.A. Thanaraj, F. Clark, and J. Muilu. Conservation of human alternative splice events in mouse. *Nucleic Acids Res*, 31(10):2544–52, May 15 2003.

T.A. Thanaraj, S. Stamm, F. Clark, J.J. Riethoven, V. Le Texier, and J. Muilu. ASD: the Alternative Splicing Database. *Nucleic Acids Res*, 32 Database issue:D64–9, Jan 1 2004.

The `FlyBase` Consortium. The `FlyBase` database of the *Drosophila* genome projects and community literature. *Nucleic Acids Res*, 31(1):172–5, Jan 1 2003.

A. Thomas and M.H. Skolnick. A probabilistic model for detecting coding regions in DNA sequences. *IMA J Math Appl Med Biol*, 11(3):149–60, 1994.

J.W. Thomas and J.W. Touchman. Vertebrate genome sequencing: building a backbone for comparative genomics. *Trends Genet*, 18(2):104–8, Feb 2002.

J.W. Thomas, J.W. Touchman, R.W. Blakesley, G.G. Bouffard, S.M. Beckstrom-Sternberg, E.H. Margulies, M. Blanchette, A.C. Siepel, P.J. Thomas, J.C. McDowell, B. Maskeri, N.F. Hansen, M.S. Schwartz, R.J. Weber, W.J. Kent, et al. Comparative analyses of multi-species sequences from targeted genomic regions. *Nature*, 424(6950):788–93, Aug 14 2003.

J.D. Tisdall. *Mastering Perl for Bioinformatics.* O'Reilly & Associates, Inc., first edition, September 2003. ISBN 0-596-00307-2.

M. Tompa. Identifying functional elements by comparative DNA sequence analysis. *Genome Res*, 11 (7):1143–4, Jul 2001.

A. Toyoda, H. Noguchi, T.D. Taylor, T. Ito, M.T. Pletcher, Y. Sakaki, R.H. Reeves, and M. Hattori. Comparative genomic sequence analysis of the human chromosome `21` Down syndrome critical region. *Genome Res*, 12(9):1323–32, Sep 2002.

E.R. Tufte. *The Visual Display of Quantitative Information.* Graphics Press USA, second edition, January 2001. ISBN 0-961-39214-2.

E.C. Uberbacher and R.J. Mural. Locating protein-coding regions in human DNA sequences by a multiple sensor-neural network approach. *Proc Natl Acad Sci U S A*, 88(24):11261–5, Dec 15 1991.

Y. Ueno, M. Arita, T. Kumagai, and K. Asai. Processing sequence annotation data using the `Lua` programming language. *Genome Inform Ser Workshop Genome Inform*, 14:154–63, 2003.

A. Ureta-Vidal, L. Ettwiller, and E. Birney. Comparative genomics: genome-wide analysis in metazoan eukaryotes. *Nat Rev Genet*, 4(4):251–62, Apr 2003.

J. Usuka and V. Brendel. Gene structure prediction by spliced alignment of genomic DNA with protein sequences: increased accuracy by differential splice site scoring. *J Mol Biol*, 297(5):1075–85, Apr 14 2000.

J.C. Venter, M.D. Adams, E.W. Myers, P.W. Li, R.J. Mural, G.G. Sutton, H.O. Smith, M. Yandell, C.A. Evans, R.A. Holt, J.D. Gocayne, P. Amanatides, R.M. Ballew, D.H. Huson, J.R. Wortman, and others (including **J.F. Abril**). The sequence of the human genome. *Science*, 291(5507):1304–51, Feb 16 2001.

Y. Wada, K. Inoue, K. Ohga, and H. Yasue. Software tool for gene mapping: `gRanch`. *Comput Appl Biosci*, 13(3):323–4, Jun 1997.

D.R. Walker and E.V. Koonin. `SEALS`: a system for easy analysis of lots of sequences. *Proc Int Conf Intell Syst Mol Biol*, 5:333–9, 1997.

S. Walsh, M. Anderson, and S.W. Cartinhour. `ACEDB`: a database for genome information. *Methods Biochem Anal*, 39:299–318, 1998.

Z. Wang, M.E. Rolish, G. Yeo, V. Tung, M. Mawson, and C.B. Burge. Systematic identification and analysis of exonic splicing silencers. *Cell*, 119(6):831–45, Dec 17 2004.

D.A. Wassarman and J.A. Steitz. Interactions of small nuclear RNA's with precursor messenger RNA during *in vitro* splicing. *Science*, 257(5078):1918–25, Sep 25 1992.

R.H. Waterston, K. Lindblad-Toh, E. Birney, J. Rogers, **J.F. Abril**, P. Agarwal, R. Agarwala, R. Ainscough, M. Alexandersson, P. An, S.E. Antonarakis, J. Attwood, R. Baertsch, J. Bailey, K. Barlow, and others (International Mouse Genome Sequencing Consortium, IMGSC). Initial sequencing and comparative analysis of the mouse genome. *Nature*, 420(6915):520–62, Dec 5 2002.

S.J. Wheelan, D.M. Church, and J.M. Ostell. Spidey: a tool for mRNA-to-genomic alignments. *Genome Res*, 11(11):1952–7, Nov 2001.

D.L. Wheeler, T. Barrett, D.A. Benson, S.H. Bryant, K. Canese, D.M. Church, M. DiCuccio, R. Edgar, S. Federhen, W. Helmberg, D.L. Kenton, O. Khovayko, D.J. Lipman, T.L. Madden, D.R. Maglott, et al. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*, 33 Database Issue:D39–45, Jan 1 2005.

D.L. Wheeler, D.M. Church, A.E. Lash, D.D. Leipe, T.L. Madden, J.U. Pontius, G.D. Schuler, L.M. Schriml, T.A. Tatusova, L. Wagner, and B.A. Rapp. Database resources of the National Center for Biotechnology Information: 2002 update. *Nucleic Acids Res*, 30(1):13–6, Jan 1 2002.

H.L. Wiegand, S. Lu, and B.R. Cullen. Exon junction complexes mediate the enhancing effect of splicing on mRNA expression. *Proc Natl Acad Sci U S A*, 100(20):11327–32, Sep 30 2003.

T. Wiehe, S. Gebauer-Jung, T. Mitchell-Olds, and R. Guigo. SGP-1: prediction and validation of homologous genes based on sequence alignments. *Genome Res*, 11(9):1574–83, Sep 2001.

T. Wiehe, R. Guigó, and W. Miller. Genome sequence comparisons: hurdles in the fast lane to functional genomics. *Brief Bioinform*, 1(4):381–8, Nov 2000.

C.L. Will, C. Schneider, A.M. MacMillan, N.F. Katopodis, G. Neubauer, M. Wilm, R. Luhrmann, and C.C. Query. A novel U2 and U11/U12 snRNP protein that associates with the pre-mRNA branch site. *EMBO J*, 20(16):4536–46, Aug 15 2001.

C.L. Will, C. Schneider, R. Reed, and R. Luhrmann. Identification of both shared and distinct proteins in the major and minor spliceosomes. *Science*, 284(5422):2003–5, Jun 18 1999.

M.D. Wilson, C. Riemer, D.W. Martindale, P. Schnupf, A.P. Boright, T.L. Cheung, D.M. Hardy, S. Schwartz, S.W. Scherer, L.C. Tsui, W. Miller, and B.F. Koop. Comparative analysis of the gene-dense *ACHE*/*TFR2* region on human chromosome 7q22 with the orthologous region on mouse chromosome 5. *Nucleic Acids Res*, 29(6):1352–65, Mar 15 2001.

V. Wood, R. Gwilliam, M.A. Rajandream, M. Lyne, R. Lyne, A. Stewart, J. Sgouros, N. Peat, J. Hayles, S. Baker, D. Basham, S. Bowman, K. Brooks, D. Brown, S. Brown, et al. The genome sequence of *Schizosaccharomyces pombe*. *Nature*, 415(6874):871–80, Feb 21 2002.

L. Woodley and J. Valcárcel. Regulation of alternative pre-mRNA splicing. *Briefings in Functional Genomics and Proteomics*, 1(3):266–77, Oct 2002.

F.A. Wright, W.J. Lemon, W.D. Zhao, R. Sears, D. Zhuo, J.P. Wang, H.Y. Yang, T. Baer, D. Stredney, J. Spitzner, A. Stutz, R. Krahe, and B. Yuan. A draft annotation and overview of the human genome. *Genome Biol*, 2(7):RESEARCH0025, 2001.

J.A. Wu and J.L. Manley. Base pairing between U2 and U6 snRNAs is necessary for splicing of a mammalian pre-mRNA. *Nature*, 352(6338):818–21, Aug 29 1991.

Q. Wu and A.R. Krainer. Splicing of a divergent subclass of AT-AC introns requires the major spliceosomal snRNAs. *RNA*, 3(6):586–601, Jun 1997.

J.R. Wyatt, E.J. Sontheimer, and J.A. Steitz. Site-specific cross-linking of mammalian U5 snRNP to the 5′ splice site before the first step of pre-mRNA splicing. *Genes Dev*, 6(12B):2542–53, Dec 1992.

Y. Xu, J.R. Einstein, R.J. Mural, M. Shah, and E.C. Uberbacher. An improved system for exon recognition and gene modeling in human DNA sequences. *Proc Int Conf Intell Syst Mol Biol*, 2:376–84, 1994a.

Y. Xu, R.J. Mural, and E.C. Uberbacher. Constructing gene models from accurately predicted exons: an application of dynamic programming. *Comput Appl Biosci*, 10(6):613–23, Dec 1994b.

Y. Xu, R.J. Mural, and E.C. Uberbacher. Inferring gene structures in genomic sequences using pattern recognition and expressed sequence tags. *Proc Int Conf Intell Syst Mol Biol*, 5:344–53, 1997.

Z. Xuan, J. Wang, and M.Q. Zhang. Computational comparison of two mouse draft genomes and the human golden path. *Genome Biol*, 4(1):R1, 2003.

J. Yang, J. Wang, Z.J. Yao, Q. Jin, Y. Shen, and R. Chen. GenomeComp: a visualization tool for microbial genome comparison. *J Microbiol Methods*, 54(3):423–6, Sep 2003.

K. Yankulov, J. Blau, T. Purton, S. Roberts, and D.L. Bentley. Transcriptional elongation by RNA polymerase II is stimulated by transactivators. *Cell*, 77(5):749–59, Jun 3 1994.

R.F. Yeh, L.P. Lim, and C.B. Burge. Computational inference of homologous gene structures in the human genome. *Genome Res*, 11(5):803–16, May 2001.

Y.T. Yu and J.A. Steitz. Site-specific crosslinking of mammalian U11 and U6atac to the 5′ splice site of an AT-AC intron. *Proc Natl Acad Sci U S A*, 94(12):6030–5, Jun 10 1997.

N. Yuhki, T. Beck, R.M. Stephens, Y. Nishigaki, K. Newmann, and S.J. O'Brien. Comparative genome organization of human, murine, and feline *MHC* class II region. *Genome Res*, 13(6A):1169–79, Jun 2003.

M.Q. Zhang. Identification of protein coding regions in the human genome by quadratic discriminant analysis. *Proc Natl Acad Sci U S A*, 94(2):565–8, Jan 21 1997.

M.Q. Zhang. Computational prediction of eukaryotic protein-coding genes. *Nat Rev Genet*, 3(9): 698–709, Sep 2002.

X.H. Zhang and L.A. Chasin. Computational definition of sequence motifs governing constitutive exon splicing. *Genes Dev*, 18(11):1241–50, Jun 1 2004.

Z. Zhang, S. Schwartz, L. Wagner, and W. Miller. A greedy algorithm for aligning DNA sequences. *J Comput Biol*, 7(1-2):203–14, Feb-Apr 2000.

D.A. Zorio and D.L. Bentley. The link between mRNA processing and transcription: communication works both ways. *Exp Cell Res*, 296(1):91–7, May 15 2004.

Index

Notes

Titles in the GBL Dissertation Series

2002-01 M. Burset.
Estudi computacional de l'especificació dels llocs d'splicing.
[Computational analysis of the splice sites definition.]
Departament de Genètica, Universitat de Barcelona.

2004-01 Sergi Castellano.
Towards the characterization of the eukaryotic selenoproteome: a computational approach.
Departament de Ciències Experimentals i de la Salut, Universitat Pompeu Fabra.

2004-02 Genís Parra.
Computational identification of genes: "ab initio" and comparative approaches.
Departament de Ciències Experimentals i de la Salut, Universitat Pompeu Fabra.

2005-01 Josep F. Abril.
Comparative Analysis of Eukaryotic Gene Sequence Features.
Departament de Ciències Experimentals i de la Salut, Universitat Pompeu Fabra.

Josep Francesc Abril Ferrando

Comparative Analysis of Eukaryotic Gene Sequence Features

Anàlisi Comparativa d'Elements de Seqüència dels Gens Eucariotes

The constantly increasing amount of available genome sequences, along with an increasing number of experimental techniques, will help to produce the complete catalog of cellular functions for different organisms, including humans. Such a catalog will define the base from which we will better understand how organisms work at the molecular level. At the same time it will shed light on which changes are associated with disease. Therefore, the raw sequence from genome sequencing projects is worthless without the complete analysis and further annotation of the genomic features that define those functions. This dissertation presents our contribution to three related aspects of gene annotation on eukaryotic genomes.

First, a comparison at sequence level of human and mouse genomes was performed by developing a semi-automatic analysis pipeline. The `SGP2` gene-finding tool was developed from procedures used in this pipeline. The concept behind `SGP2` is that similarity regions obtained by `TBLASTX` are used to increase the score of exons predicted by `geneid`, in order to produce a more accurate set of gene structures. `SGP2` provides a specificity that is high enough for its predictions to be experimentally verified by RT-PCR. The RT-PCR validation of predicted splice junctions also serves as example of how combined computational and experimental approaches will yield the best results.

Then, we performed a descriptive analysis at sequence level of the splice site signals from a reliable set of orthologous genes for human, mouse, rat and chicken. We have explored the differences at nucleotide sequence level between U2 and U12 for the set of orthologous introns derived from those genes. We found that orthologous splice signals between human and rodents and within rodents are more conserved than unrelated splice sites. However, additional conservation can be explained mostly by background intron conservation. Additional conservation over background is detectable in orthologous mammalian and chicken splice sites. Our results also indicate that the U2 and U12 intron classes have evolved independently since the split of mammals and birds. We found neither convincing case of interconversion between these two classes in our sets of orthologous introns, nor any single case of switching between AT-AC and GT-AG subtypes within U12 introns. In contrast, switching between GT-AG and GC-AG U2 subtypes does not appear to be unusual.

Finally, we implemented visualization tools to integrate annotation features for gene-finding and comparative analyses. One of those tools, `gff2ps`, was used to draw the whole genome maps for human, fruitfly and mosquito. `gff2aplot` and the accompanying parsers facilitate the task of integrating sequence annotations with the output of homology-based tools, like `BLAST`. We have also adapted the concept of pictograms to the comparative analysis of orthologous splice sites, by developing `compi`.

GBL Dissertation Series
Universitat Pompeu Fabra

ISBN XX-XXX-XXXX-X