Capítol 5.

Pmut: Un servidor d'internet per la predicció de mutacions puntuals.

[*Aquesta pàgina ha estat deixada en blanc intencionadament*]

INTRODUCCIÓ

La biomedicina necessita desenvolupar eines ràpides i eficients per l'anotació de la gran quantitat de dades aportades pels projectes actuals de seqüenciació (Lander et al., 2001; Venter et al., 2001). En particular és imprescindible tenir programes per l'anotació de les mutacions puntuals o SNPs, que explicarien prop del 90% de la variació intraespecífica (Cargill et al., 1999; Chakravarti, 2001). Ja en el primer esborrany del projecte genoma humà apareixen fins a 1,4 milions de SNPs (Sachidanandam et al., 2001) i s'avalua que la freqüència de SNPs per individu és de 1 cada 1000 nucleòtids (Sunyaev et al., 2000). D'aquesta variabilitat alguns SNPs més rars serien l'origen de les mutacions responsables de malalties hereditàries monogèniques, mentre que d'altres SNPs més freqüents serien responsables de malalties més complexes i comunes, que són les malalties poligèniques (Collins et al., 1998) i finalment un bon número de SNPs serien irrelevants per a la salut de l'organisme. En els darrers anys s'han presentat diferents programes que incideixen en la caracterització de SNPs causants de patologia així com la seva predicció (Chasman & Adams, 2001; Ng & Henikoff, 2002; Saunders & Baker, 2002; Sunyaev et al., 2001; Wang & Moult, 2003). En aquest capítol es presenta el servidor d'internet anomenat Pmut (www.mmb2.pcb.ub.es:8080/PMut/PMut.jsp) (veure figura 1) que implementa, per a l'ús públic, la metodologia desenvolupada en el nostre grup per a la predicció de mutacions puntuals en proteïnes (Ferrer-Costa et al., 2002; Ferrer-Costa et al., 2004). La metodologia es basa exclusivament en propietats derivades de la seqüència, fet que incrementa el seu rang d'aplicació pràctica.

MATERIALS I MÈTODES

Estructuració del servidor

Pmut és un servidor per la predicció del caràcter patològic de les mutacions puntuals. Treballa en dos nivells diferents: i) retorna informació d'una base de dades local de "punts calents mutacionals" i ii) analitza mutacions puntuals en una proteïna problema. Els resultats són mostrats en diferents fitxers de text, i si existeix l'estructura també de manera gràfica en dues i tres dimensions.

Predictor Pmut

Per tal d'obtenir prediccions de PMUT l'usuari necessita d'introduir la seqüència de la proteïna (en codi d'una sola lletra) o alternativament el codi de la proteïna a Swissprot/trEMBL. Posteriorment, ha de seleccionar la posició de la mutació i decidir si analitza una sola mutació (per defecte) o fer una anàlisi exhaustiva canviant l'aminoàcid original pels 19 restants o al subconjunt d'aminoàcids accessible per mutació d'un únic nucleòtid. Alternativament, pot triar d'analitzar la seqüència completa (*Mutation Hot-Spot analysis*). El programa recull de les pròpies bases de dades (Ferrer-Costa et al., 2002; Ferrer-Costa et al., 2004) una sèrie de paràmetres per la descripció de la mutació en estudi, com són valors de matrius de mutació, paràmetres de volums d'aminoàcid, propensions d'estructura secundària, descriptors d'hidrofobicitats i potencials de seqüència (Ferrer-Costa et al., 2002; Ferrer-Costa et al., 2004). Depenent de la xarxa neural usada (NN) (veure més avall) altres paràmetres relacionats amb l'estructura (accessibilitat i estructura secundària predites) són derivats pel programa PHD (Rost & Sander, 1993) a partir de la seqüència . Finalment altres descriptors (Ferrer-Costa et al., 2004) són obtinguts d'alineaments múltiples de seqüència que es generen automàticament a partir dels resultats de córrer PSI-Blast (Altschul et al., 1997) a dues iteracions sobre una versió no redundant de la base de dades

SwissProt/trEMBL.

Alternativament, l'usuari pot introduir ell mateix l'alineament múltiple de seqüències obtingut per altres fonts (per exemple, alineaments de la base de dades Pfam (Bateman et al., 2002)). Els alineaments estan limitats a 300 seqüències (E-values per sota 10^{-4}). L'usuari és advertit sobre la qualitat dels resultats quan el programa reconeix un alineament múltiple de Blast pobre. El programa deriva diferents paràmetres d'aquests alineaments múltiples com un índex de variabilitat, l'entropia de seqüència i índexs de PSSM.

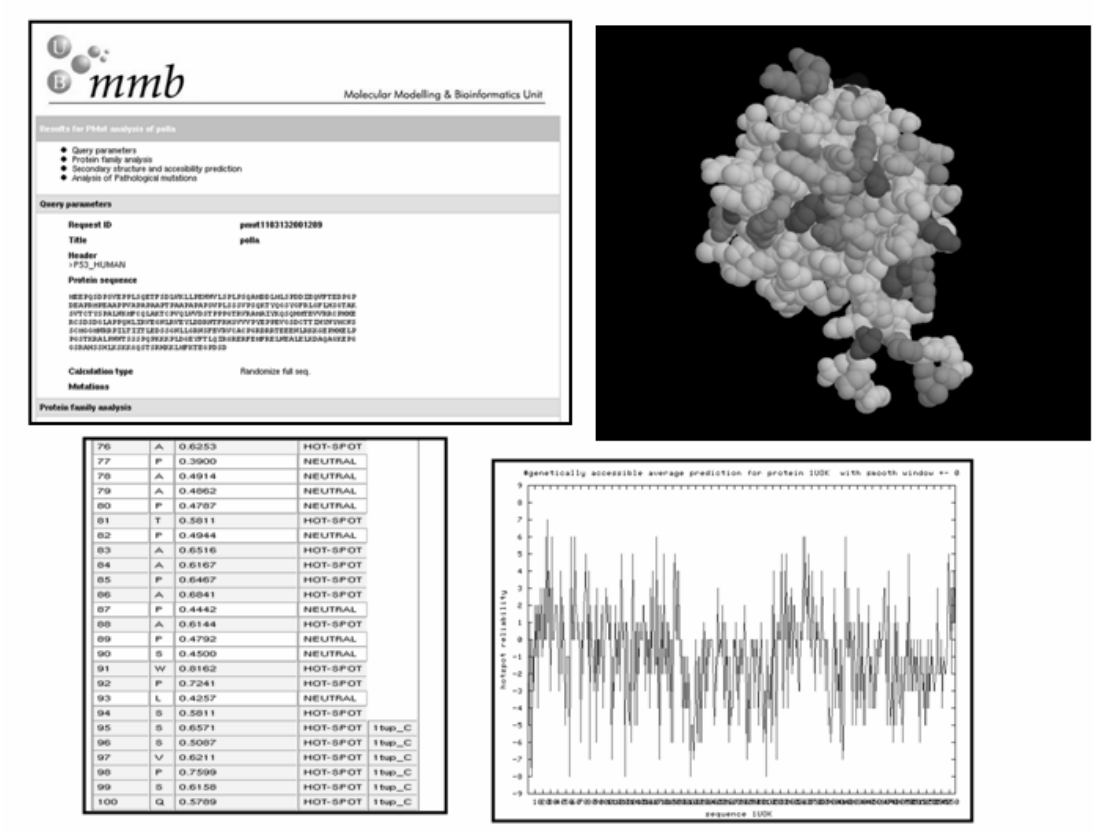

Fig 2. Esquema representatiu de les dades que retorna el servidor i els seus diferents formats (taules, gràfiques, representacions moleculars…)

S'han implementat dues xarxes neurals (NN) com a eines predictives: una de gran amb una capa oculta (20 nodes) i 5 descriptors (Ferrer-Costa et al., 2002; Ferrer-Costa et al., 2004) i una de petita (20 nodes, sense capa oculta) amb 3

paràmetres (Ferrer-Costa et al. en preparació). Les dues xarxes van ser entrenades usant validació creuada amb dades humanes (Ferrer-Costa et al., 2002; Ferrer-Costa et al., 2004) (Ferrer-Costa et al en preparació). La xarxa gran és la opció determinada per defecte, però el model simple pot ser una bona alternativa quan PMUT és usat per predir DAMUs en organismes distants evolutivament. El resultat final del programa és: i) un índex de caràcter patològic de 0 a 1 (mutacions associades amb un índex per sobre 0.5 són predites com a patològiques) i ii) un índex que varia de 0 (baixa) a 9 (alta) que correspon al nivell de fiabilitat de la predicció. Addicionalment, el programa permet a l'usuari de recuperar tota la informació intermèdia (alineaments, resultats de Blast, PHD...).

El servidor permet mostrar el lloc de la mutació en l'estructura de la proteïna, quan aquesta existeix, usant un codi de color per tal de traçar la patogenicitat associada amb la mutació. Per aquest proposit, s'ha usat BLAST (Altschul et al., 1997) per trobar seqüències altament homòlogues en la base de dades PDB (Berman et al., 2000). Totes les proteïnes amb una identitat de seqüència inferior al 70% i totes aquelles de més de 200 residus però amb menys de 70 residus alineats són eliminades. L'alineament permet mapar la mutació en l'estructura tridimensional de la proteïna (o d'un homòleg proper). Un *script* de Rasmol (Sayle & Milner-White, 1995) és creat per tal de mostrar l'estructura tridimensional de la proteïna amb el lloc de la mutació representat en un codi de color diferent que permet la seva identificació. L'estructura es pot visualitzar de manera remota usant el *plug-in* CHIME (MDL Information Systems, Inc) o alternativament l'arxiu es pot guardar localment per la seva inspecció usant RASMOL (Sayle & Milner-White, 1995).

La base de dades pmut

S'han precalculat els perfils de mutació per totes les proteïnes del clúster 90 de la base de dades PDB (Berman et al., 2000) (conté un representant per cada clúster

que comparteixen un 90% d'identitat de seqüència). Amb aquest propòsit s'han mutat tots els residus de cada proteïna a tots els possibles 19 altres aminoàcids calculant N*19 matrius de patogenicitat (on N és el número de residus de la proteïna). Aquesta matriu és manipulada per definir punts calents de mutació en diferents maneres: i) el màxim, la mitja i el mínim índex de patogenicitat en cada residu de la proteïna, ii) l'índex de patogenicitat associat amb la mutació a Ala (*alanine-scanning*) de tots els residus, iii) el màxim, la mitja i el mínim índex de patogenicitat associat a les mutacions genèticament accessibles (p.e. aquelles que impliquen només un canvi d'un sol nucleòtid) en cada posició de la proteïna.
Aquesta informació es pot recollir del servidor (veure figures 1 i 2) en diferents maneres: i) com a fitxer de text, ii) com a gràfiques bidimensionals dels perfils de mutació, o iii) en forma gràfica, on cada posició de la proteïna és mostrada d'acord amb cada índex patològic.
L'usuari és avisat quan la proteïna analitzada no és humana.

Implementació del programari i el seu ús

Pmut és accessible de manera lliure a través d'una interfície Web al lloc d'internet del grup Molecular Modelling & Bioinformatics (http://mmb2.pcb.ub.es:8080/PMut). És escrit com un *servlet* de Java. Pmut té un nucli escrit en C, complementat per una sèrie de *scripts* de Perl responsables del funcionament de tots els serveis incloent els programes auxiliars (veure figura 1). Els gràfics 2D són obtinguts usant el programari GNUPLOT (http://www.gnuplot.info/) que corre en el servidor i donant fitxers d'imatge estàndard. Els resultats 3D són gestionats amb *scripts* de Rasmol, la seva visualització requereix l'ús tant de Rasmol (Sayle & Milner-White, 1995) o el *plug-in* CHIME (MDL Information Systems) en la banda del client. Els càlculs corren usant una cua en batch en el servidor i l'usuari és avisat del seu acabament ja sigui des de la pròpia pàgina Web o bé per correu electrònic. Un càlcul típic dura 1-10 minuts depenent de la càrrega del servidor i dels paràmetres del càlcul.

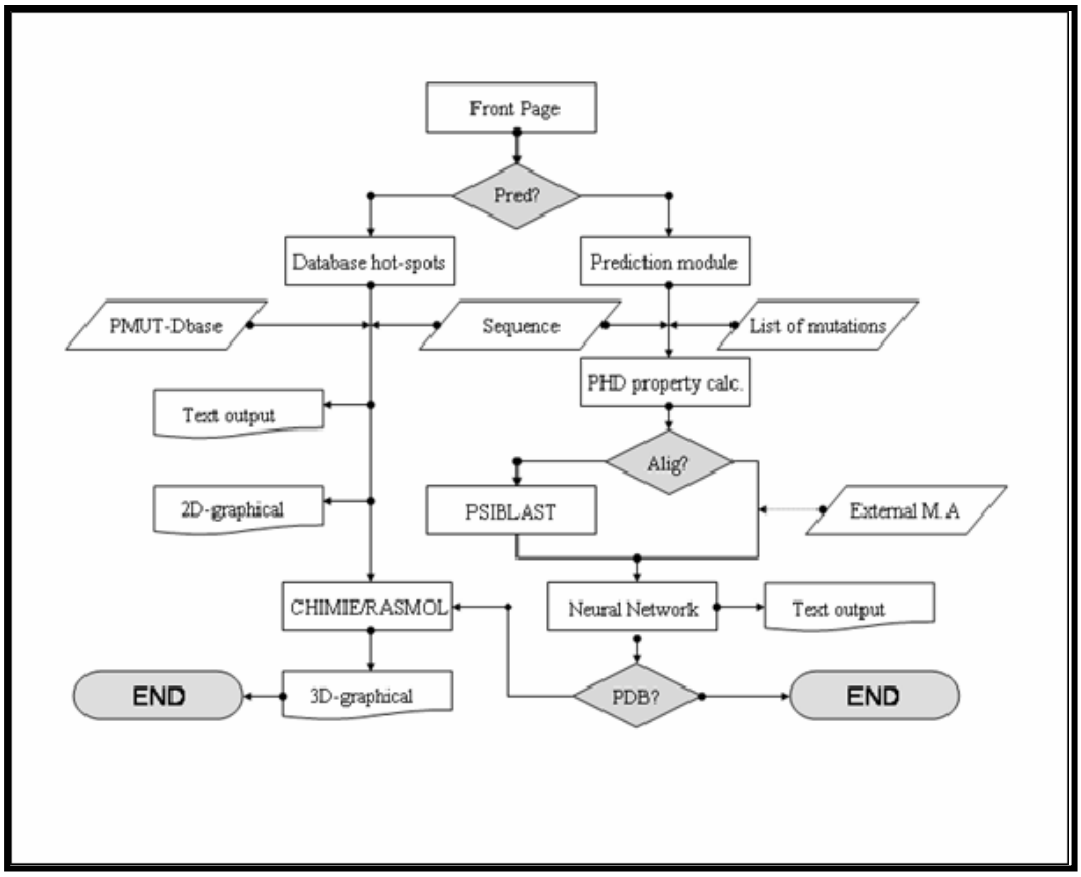

Fig2. Diagrama de Flux que defineix el servidor Pmut i la seva base de dades.

Una versió limitada del predictor-PMUT amb l'anàlisi de Hot-Spot és també accessible via servei de Web corrent segons el BioMoby (http://www.biomoby.org) estàndard (http://www.inab.org).

El servidor PMUT ha estat implementat també com a part del servidor PupasView. Aquest servidor ha estat desenvolupat al grup del professor Dopazo amb la intenció de crear una eina accessible via web per l'accés a la selecció de SNPs amb potencial efecte fenotípic. El servidor PupasView constitueix un entorn interactiu on la informació funcional i de freqüència poblacional pot ser usada com a filtres seqüencials sobre els paràmetres de desequilibri de lligament per obtenir un llista final de SNPs òptims per propòsits de genotipat. PupasView

és el primer servidor que intregra efectes fenotípics causats per SNPs a nivell traduccional com transcripcional. El servidor selecciona SNPs que afecten a regions conservades que la maquinària cel·lular usa pel correcte processat dels gens (regions flanquejants intró/exó o *enhancers* de *splicing* exònic), regions predites d'unió a factors de transcripció (TFBS *predicted transcription factor binding sites*), i canvis en aminoàcids que són predits com a patològics. El servidor PupasView és accessible via web a les següents adreces. http://pupasview.bioinfo.cnio.es i http://www.pupasnp.org.

[*Aquesta pàgina ha estat deixada en blanc intencionadament*]

III. ARTICLE DE RECERCA

PMUT: a web-based tool for the annotation of pathological mutations on proteins. Carles Ferrer-Costa, Josep Lluís Gelpí, Leire Zamakola, Ivan Parraga, Xavier de la Cruz, Modesto Orozco. Bioinformatics doi: 10.1093/bioinformatics/bti486.

[*Aquesta pàgina ha estat deixada en blanc intencionadament*]

BIOINFORMATICS APPLICATIONS NOTE Vol. 00 no. 0 2005, pages 1–3
doi:10.1093/bioinformatics/bti486

Structural bioinformatics

PMUT: a web-based tool for the annotation of pathological mutations on proteins

Carles Ferrer-Costa[1], Josep Lluis Gelpí[1,2,*], Leire Zamakola[1,3], Ivan Parraga[1,3], Xavier de la Cruz[1,4] and Modesto Orozco[1,2,3,*]

[1]Molecular Modeling and Bioinformatics Unit, Institut de Recerca Biomédica, Parc Científic de Barcelona, Josep Samitier 1-5, Barcelona 08028, Spain, [2]Departament de Bioquímica i Biología Molecular, Facultat de Química, Universitat de Barcelona, Martí i Franquès 1, Barcelona 08028, Spain, [3]Instituto Nacional de Bioinformàtica, Parc Científic de Barcelona, Josep Samitier 1-5, Barcelona 08028, Spain and [4]Institució Catalana per la Recerca i Estudis Avançats (ICREA), Passeig Lluís Companys 23, 08018 Barcelona, Spain

Received on March 8, 2005; revised on April 13, 2005; accepted on May 3, 2005
Advance Access publication ...

ABSTRACT

Summary: PMUT allows the fast and accurate prediction (~80% success rate in humans) of the pathological character of single point amino acidic mutations based on the use of neural networks. The program also allows the fast scanning of mutational hot spots, which are obtained by three procedures: (1) alanine scanning, (2) massive mutation and (3) genetically accessible mutations. A graphical interface for Protein Data Bank (PDB) structures, when available, and a database containing hot spot profiles for all non-redundant PDB structures are also accessible from the PMUT server.

Availability: PMUT is freely accessible at http://mmb.pcb.ub.es:8080/PMut/PMut.jsp

Contact: modesto@mmb.pcb.ub.es

Supplementary information:

INTRODUCTION

The processing of the massive amount of data on single nucleotide polymorphisms (SNPs) requires the development of automatic annotation tools to determine the potential pathological character of a given SNP. Some of these programs (e.g. http://pupasnp.bioinfo.cnio.es/) trace the positioning of SNPs in the genome, detecting when they occur in a functionally important region. Others are centered on the study of non-synonymous mutations mapped on proteins; for discussions see (Chasman and Adams, 2001; Ferrer-Costa *et al.*, 2002, 2004, 2005; Ng and Henikoff, 2002; Saunders and Baker, 2002; Sunyaev *et al.*, 2001; Wang and Moult, 2001).

Our group has developed an accurate (>80% success rate) and robust methodology to predict disease-associated mutations (DAMUs) (Ferrer-Costa *et al.*, 2002, 2004, 2005). The method is based on the use of neural networks (NNs) trained with a large database of neutral mutations (NEMUs) and pathological mutations. In this paper we present the PMUT server, which implements our predictive models and complementary tools that can help in the annotation of SNPs.

*To whom correspondence should be addressed.

SERVER STRUCTURE

PMUT works at two different levels (Fig. 1S): (1) it retrieves information from a local database of mutational hotspots and (2) it analyzes a given SNP in a specific protein. Results are displayed in the form of various text files and, when the structure is experimentally known, 2-D and 3-D plots are also available.

PMUT PREDICTOR

The first input to PMUT is either the sequence of the protein or its SwissProt/trEMBL code. The user has to select the mutation site and whether to analyze a single mutation (default) or to perform a complete mutation scan at this position. The program can simulate massive single-point mutation along the whole sequence (Mutation Hot-Spot analysis), helping to detect regions where mutations are expected to have a large pathological impact. Irrespective of the selection, the program retrieves a series of parameters describing the mutation (Ferrer-Costa *et al.*, 2002, 2004) from (1) its internal databases, (2) PHD output (Rost and Sander, 1993) and (3) multiple alignments. The latter are either introduced by the user [e.g. from the PFAM database (Bateman *et al.*, 2002)] or automatically generated by the program from a two-iterations PSI-Blast (Altschul *et al.*, 1997) run on a non-redundant SwissProt/trEMBL database.

Two NNs are implemented as predictor engines: a large one (the default) with 1 hidden layer, 20 nodes and 15 descriptors (Ferrer-Costa *et al.*, 2002, 2004) and a small one (20 nodes, no hidden layer) with 3 parameters (Ferrer-Costa *et al.* submitted for publication). Both NNs were carefully trained with human mutational data. The final output is always (1) a pathogenicity index ranging from 0 to 1 (indexes >0.5 signal pathological mutations) and (2) a confidence index ranging from 0 (low) to 9 (high). Additionally, the program allows the user to retrieve all the intermediate information (alignments, Blast and PHD outputs, etc.) used in PMUT predictions.

The PMUT server allows the display of the mutation site on the protein structure (when this is available) using a color code to trace the pathogenicity associated with the mutation. For this purpose, we used Blast (Altschul *et al.*, 1997) to find highly homologous

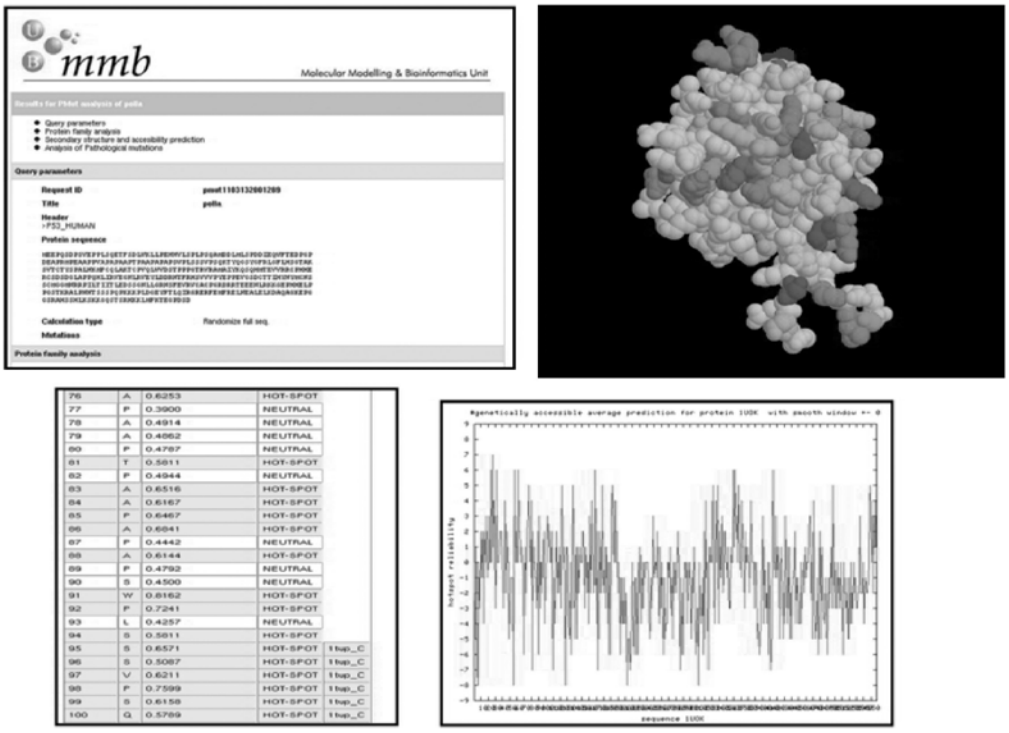

Fig. 1. Examples of outputs obtained from the PMUT database.

sequences in the complete PDB (Berman *et al.*, 2000). All proteins with <70% local identity and those with >200 residues and covered by PDB with <70% identical were eliminated. A Rasmol (Sayle and Milner-White, 1995) script is automatically created to display the mutation site on the protein structure. Visualization can be done remotely using the Chime plug-in (MDL Information Systems, Inc.) or alternatively the file can be download for local inspection using Rasmol.

PMUT DATABASE

We have pre-computed the mutation profiles of all the proteins in the 90% identity cluster of the PDB database (Berman *et al.*, 2000). For this purpose we mutated all the residues of each protein to all 19 possible alternative amino acids. The mutation matrix is manipulated to define mutation hot spots in different ways: (1) maximum, mean and minimum pathogenicity indexes in each mutation site, (2) the pathogenicity index associated with the mutation to Ala (alanine-scanning) of all the residues and (3) the maximum, mean and minimum pathogenicity indexes associated with the genetically accessible mutations (i.e. those implying only one nucleotide change) in each position of the protein.

All this information can be retrieved from the server (Fig. 1) in text and graphical formats (Fig. 1). To avoid over-interpretation of the results, the user is alerted when the protein is not human. The help section includes a description of the validity of prediction in non-human proteins using human-trained NNs.

SOFTWARE IMPLEMENTATION AND USE

PMUT is freely accesible through a web interface at the Molecular Modeling and Bioinformatics website (http://mmb.pcb.ub.es:8080/PMut/PMut.jsp). The interface is written as a Java servlet. PMUT has a core written in C, complemented with a series of Perl scripts responsible for the overall workflow, including the execution of auxiliary programs. Two-dimensional graphical output is obtained using the GNUPLOT software (http://www.gnuplot.info/) running on the server and is provided as standard image files. The 3-D outputs are provided as Rasmol scripts, and visualization requires the use of Rasmol or the Chime plug-in on the client side. Calculations are run using a batch queue in the server and the user is informed of their completion either from the web page or by email. A limited version of PMUT Predictor providing a hot spot analysis is also available as a web service running according to the BioMoby standard (http://www.biomoby.org; http://www.inab.org).

ACKNOWLEDGEMENTS

This work has been supported by the Instituto Nacional de Bioinformática (INB-Genóma España), Fundación Ramón-Areces and the

Spanish Ministry of Education and Science (BIO2003-06848 and GEN2001-4758-C07-07).

REFERENCES

Altschul,S.F. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

Bateman,A. *et al.* (2002) The Pfam protein families database. *Nucleic Acids Res.*, **30**, 276–280.

Berman,H.M. *et al.* (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.

Chasman,D. and Adams,R.M. (2001) Predicting the functional consequences of non-synonymous single nucleotide polymorphisms: structure-based assessment of amino acid variation. *J. Mol. Biol.*, **307**, 683–706.

Ferrer-Costa,C. *et al.* (2002) Characterization of disease-associated single amino acid polymorphisms in terms of sequence and structure properties. *J. Mol. Biol.*, **315**, 771–786.

Ferrer-Costa,C. *et al.* (2004) Sequence-based prediction of pathological mutations. *Proteins*, **57**, 811–819.

Ferrer-Costa,C. *et al.* (2005) Use of bioinformatics tools for the annotation of disease-associated mutations in animal models. *Proteins*, in press.

Ng,P.C. and Henikoff,S. (2002) Accounting for human polymorphisms predicted to affect protein function. *Genome Res.*, **12**, 436–446.

Rost,B. and Sander,C. (1993) Prediction of protein secondary structure at better than 70% accuracy. *J. Mol. Biol.*, **232**, 584–599.

Saunders,C.T. and Baker,D. (2002) Evaluation of structural and evolutionary contributions to deleterious mutation prediction. *J. Mol. Biol.*, **322**, 891–901.

Sayle,R.A. and Milner-White,E.J. (1995) RASMOL: biomolecular graphics for all. *Trends Biochem. Sci.*, **20**, 374.

Sunyaev,S. *et al.* (2001) Prediction of deleterious human alleles. *Hum. Mol. Genet.*, **10**, 591–597.

Wang,Z. and Moult,J. (2001) SNPs, protein structure, and disease. *Hum. Mutat.*, **17**, 263–270.

[*Aquesta pàgina ha estat deixada en blanc intencionadament*]

IV. ARTICLE DE RECERCA

PUPASVIEW: A visual tool for selecting suitable SNPs, with putative pathological effect in genes, for genotyping purpose. Lucia Conde, Juan M. Vaquerizas, Carles Ferrer-Costa, Xavier de la Cruz, Modesto Orozco and Joaquín Dopazo. Nucleic Acids Research doi: 10.1093/nar/gki476.

[*Aquesta pàgina ha estat deixada en blanc intencionadament*]

Nucleic Acids Research, 2005, Vol. 33, Web Server issue **W1–W5**
doi:10.1093/nar/gki476

PupasView: a visual tool for selecting suitable SNPs, with putative pathological effect in genes, for genotyping purposes

Lucía Conde[1], Juan M. Vaquerizas[1], Carles Ferrer-Costa[2], Xavier de la Cruz[2,4], Modesto Orozco[2,3,5] and Joaquín Dopazo[1,6,*]

[1]Bioinformatics Unit, Centro Nacional de Investigaciones Oncológicas (CNIO), Madrid 28029, Spain, [2]Molecular Modelling and Bioinformatics Unit, Institut de Recerca Biomèdica and [3]Structure and Modelling Node INB, Parc Científic de Barcelona, Barcelona 08028, Spain, [4]Institució Catalana per la Recerca i Estudis Avançats (ICREA), 08018 Barcelona, Spain, [5]Departament de Bioquímica i Biología Molecular Facultat de Química, Universitat de Barcelona, Barcelona 08028, Spain and [6]Functional Genomics Node, INB, CIPF Valencia 46013, Spain

Received February 14, 2005; Revised and Accepted April 15, 2005

ABSTRACT

We have developed a web tool, PupasView, for the selection of single nucleotide polymorphisms (SNPs) with potential phenotypic effect. PupasView constitutes an interactive environment in which functional information and population frequency data can be used as sequential filters over linkage disequilibrium parameters to obtain a final list of SNPs optimal for genotyping purposes. PupasView is the first resource that integrates phenotypic effects caused by SNPs at both the translational and the transcriptional level. PupasView retrieves SNPs that could affect conserved regions that the cellular machinery uses for the correct processing of genes (intron/exon boundaries or exonic splicing enhancers), predicted transcription factor binding sites and changes in amino acids in the proteins for which a putative pathological effect is calculated. The program uses the mapping of SNPs in the genome provided by Ensembl. PupasView will be of much help in studies of multifactorial disorders, where the use of functional SNPs will increase the sensitivity of the identification of the genes responsible for the disease. The PupasView web interface is accessible through http://pupasview.bioinfo.cnio.es and through http://www.pupasnp.org.

INTRODUCTION

Single nucleotide polymorphisms (SNPs) are the simplest and most frequent type of DNA sequence variation among individuals and, with the recent availability of high-throughput methodologies, are considered one of the most powerful tools in the search for e.g. disease susceptibility genes and drug response-determining genes (1,2). However, complex diseases, for which markers display weak associations, still constitute a challenge. Most probably, advancement in the knowledge of such diseases will come from improved genotyping methods in combination with the proper bioinformatics design strategies (3).

It is generally believed that multigenicity reflects disruptions in proteins that participate in a protein complex or in a pathway (4). Typically, SNPs have been used as markers; that is, the real determinant of the disease was not the SNP itself but some other mutation in linkage disequilibrium (LD) with it. Because of this, the use of functional SNPs could be an important factor in increasing significantly the sensitivity of association tests. In fact, several complex genetic disorders such as Alzheimer's disease (5) and Crohn' disease (6) have been associated with functional SNPs, lending weight to strategies giving priority to candidate markers based upon predictable function. Several estimations suggest that, on average, some 20% of SNPs could directly damage proteins (7).

Much attention has been focused on modelling by different methods the possible phenotypic effect of SNPs that cause amino acid changes (7–13), and only recently has interest focused on functional SNPs affecting regulatory regions or the splicing process (14). However, there is increasing evidence that many human disease genes are the result of exonic or non-coding mutations affecting regulatory regions (15–17). A recent large-scale screening over a set of 16 chromosomes found SNPs in the promoter regions of 35% of the genes, and experimental evidence suggested that around a third of

*To whom correspondence should be addressed. Email: jdopazo@ochoa.fib.es
Present address:
Lucia Conde, Juan M. Vaquerizas and Joaquín Dopazo, Department of Bioinformatics, Centro de Investigación Príncipe Felipe, Valencia 46013, Spain

W2 Nucleic Acids Research, 2005, Vol. 33, Web Server issue

promoter variants may alter gene expression to a functionally relevant extent (18). Alternative splicing produced by mutations in intron/exon junctions, or in distinct binding motifs, such as exonic splicing enhancers (ESEs) (19), has also been related to different diseases (20). In fact, it has been estimated that 15% of point mutations that result in human genetic diseases cause RNA splicing defects (21).

In addition to functional information, population frequency is another important factor to be taken into account when selecting SNPs. Thus, infrequent polymorphisms will be of scarce interest as markers. Also, LD is another interesting factor in selecting SNPs as markers since, if two SNPs are in strong LD, only one of them will provide enough information for any association or linkage test.

With the idea of selecting optimal sets of SNPs using as much information as possible on putative phenotypic effect, population frequencies and LD, we have developed PupasView (Putative Phenotypic Alterations caused by SNPs Viewer), a server that can be used alone or in combination with PupaSNP (14).

PupasView works not only as a viewer of where SNPs are located, but also as a selector in which different filters based on combinations of functionality and population frequencies can be interactively applied over the LD parameters in order to obtain an optimal selection of SNPs for genotyping studies, in such a way that with a minimum number of SNPs maximum information on the genic region is obtained.

Criteria to consider an SNP a good candidate for genotyping studies

There are three important properties for an SNP to be considered an optimal candidate for genotyping purposes: functional effect, minor allele frequency and LD with respect to other SNPs. Finding such optimal SNPs is not always possible, but the idea behind PupasView is to facilitate the selection process in order to achieve a final collection of SNPs bearing the maximum amount of information. PupasView works as an SNP selector. Different filters can be interactively applied to the LD information available based on distinct functional properties, cross-species conservation and population frequency. This permits a final selection of a minimum number of SNPs with optimal properties in terms of population frequencies and potential phenotypic effect.

Finding SNPs with potential phenotypic effect

PupasView uses a precompiled database which contains a collection of dbSNP entries mapped to the Golden Path genome assembly, as implemented in the human section of Ensembl (http://www.ensembl.org). Part of this database is common to the PupaSNP program (14). The SNPs have been labelled according to their potential effects on the phenotype. We have taken into account both transcriptional and gene product levels. Regions 10 000 bp upstream of the genes belonging to the promoter region of each gene in the list have been scanned for the presence of possible different regulatory motifs. These include alterations in:

(i) *Transcription factor binding sites*. Promoter regions were scanned for the presence of possible transcription factor binding sites. The program Match (22) was used for this purpose, using only high-quality matrices and with a cut-off to minimize false positives from the Transfac database (23). SNPs located within these motifs are considered to have a putative phenotypic effect in the expression of the gene. Almost four million such motifs were found, with 130 373 SNPs mapping onto them.

(ii) *Intron/exon border consensus sequences*. Ensembl APIs (24) were used to extract the intron/exon organization of the genes and the corresponding sequences. The two conserved nucleotides on each side of the splicing point, which constitute the splicing signal (21), were then located and all the SNPs altering these signals were recorded. More than 700 000 intron/exon boundaries could be defined in human genes with 1786 SNPs mapping onto them.

(iii) *ESEs*. Mutations that inactivate or activate an ESE sequence may result in exon skipping, errors in alternative splicing patterns, malformation and so on. Different classes of ESE consensus motifs have been described, but they are not always easily identified. Exon sequences were scanned to identify putative ESEs responsive to the human SR proteins SF2/ASF, SC35, SRp40 and SRp55, using the available weight matrices (20). A score was obtained that is related to the likelihood that the site found is a real ESE. Only ESE sites with scores over the threshold [see (20) for details] were taken into account in the analysis. More than 11 million ESEs were found, with 299 106 SNPs located in them.

(iv) *Triplex-forming oligonucleotide target sequences* (TTSs). It has been found that the population of TTSs is much more numerous than expected from simple random models (25). The population of TTSs is large in the whole genome, without major differences between chromosomes, but with a large concentration in regulatory regions, especially in promoter zones, which suggests a tremendous potential for triplex strategy in the control of gene expression (25). Although the role of TTSs in regulation is still a matter of speculation, the program also reports SNPs disrupting these structures. Some 5.4 million putative triplex-forming sequences were found, and 364 314 SNPs mapped onto them.

(v) *SNPs in exons that cause an amino acid change*. Any SNP causing a change of amino acid, independent of any speculation on its possible phenotypic effect, is reported. There are 45 906 such SNPs.

(vi) *SNPs in exons that cause an amino acid change with putative pathological effect*. The putative pathological effect of an amino acid change can be predicted using neural networks (NNs) carefully trained to predict disease-associated amino acidic polymorphism (12,13). The server implements a small NN (1 hidden layer and 20 nodes) and three sequence-derived descriptors (PAM40, PSSM and variability), which are either retrieved from databases or determined internally from multiple alignments using two-iterations PSI-Blast (26) run over a non-redundant SwissProt/TrEMBL database. The trained method displays a success rate >80% in cross-validation experiments. According to the algorithm, 19 309 SNPs displayed a high probability of having pathological effect.

(vii) *Human–mouse conserved regions*. Untranslated whole genome comparisons by BLASTZ were performed for species pairs which are thought to be similar enough to be able to detect homology directly at the DNA level (27).

Of particular interest is mouse (or rat) because of its phylogenetic position with respect to humans: distant enough to interpret conservation as important but not so distant as to lose most of the similarity. The phenotypic effect of a change in such regions is quite speculative, but cross-species conservation can be useful in cases in which no other information is available. It is also useful for reinforcing the likelihood of other predictions (e.g. an ESE in a conserved region is more likely to be real than one in a non-conserved region).

Frequency information and validation status

There are >10 million SNPs stored in the last build of dbSNP (build 124), and more than half of these have been validated by different means (http://www.ncbi.nlm.nih.gov/SNP/snp_summary.cgi). Validation status is annotated and is an important field in terms of trusting an SNP. But, in addition to being real, an SNP must exist in the population at frequencies which make it a suitable marker. Very infrequent SNPs are not suitable for association or linkage studies. For almost half a million SNPs frequency data in different populations are available.

Blocks and LD parameters

LD measures the correlation between two neighbouring genetic variants in a specific population. The program HaploView (28) is used to infer blocks using different procedures. In one of the most common procedures (29), 95% confidence bounds based on the D' LD parameter are generated and each comparison is called 'strong LD', 'inconclusive' or 'strong recombination'. A block is created if 95% of informative (i.e. non-inconclusive) comparisons are 'strong LD'. A block can be considered a region with a low recombination rate. Ideally, a block could properly be described by a unique SNP. Two other methods are used: the four gamete rule (30) and the Solid Spine of LD (28). Blocks are displayed in the bottom of the PupasView window. Also D', R^2 and LOD parameters between adjacent SNPs can be visualized by placing the cursor between them. Only HapMap genotyped SNPs (31) are used to calculate blocks and LD parameters.

The web interface of the SNPs selector

The main purpose of PupasView is to provide the user with an optimal set of SNPs for genotyping experiments by filtering the annotated SNPs using a series of filters related to their impact in protein functionality and pathology, their population frequency and LD.

The input is a gene identifier (Ensembl IDs or external IDs, which include GenBank, Swissprot/TrEMBL and other gene IDs supported by Ensembl). The program can also be invoked from PupaSNP. The program presents a list of options that can be selected and applied as many times as desired. The options include

- Validation status obtained from dbSNP
- Type of SNP (coding, intron, untranslated region, local), according to its position in the gene
- Frequency and population, an option that allows the possibility of filtering by a range of frequencies of the minor allele in one or more populations (Europe; Europe, multinational; Europe, North America; North America; Central/South America; North/East Africa and Middle East; Central/South Africa; West Africa; Central Asia; East Asia; Pacific; multinational; unknown; HapMap)
- Functional properties as follows:
  - non-synonymous SNPs [all or only those predicted as pathological by the pmut algorithm (12,13)]
  - SNPs disrupting predicted transcription factor binding sites (all or only those that are in regions conserved in the mouse genome)
  - SNPs disrupting predicted ESEs (all or only those that are in regions conserved in the mouse genome)
  - SNPs disrupting potential triplex-forming regions (all or only those that are in regions conserved in the mouse genome)
  - SNPs disrupting intron/exon boundaries
  - regions conserved in mouse
- Options for the way in which blocks are constructed:
  - confidence intervals (29)
  - four gamete rule (30)
  - Solid Spine of LD (28).

Figure 1 shows the view of the results. The viewer of PupasView has been constructed using Ensembl APIs (24). Figure 1A shows the result of running PupasView on the gene TP53 without applying any filter. All the SNPs in the gene and the neighbourhood are displayed. If the cursor is over an SNP, information on it is displayed by means of pop-up text. Figure 1B shows a subselection of these SNPs obtained after selecting only SNPs for which population frequency was available. Finally, Figure 1C shows the selection obtained if only SNPs with putative functional effect are chosen. This will constitute the final, reduced subset of optimal SNPs. The upper horizontal bar below the figure represents LD parameters (which can be individually obtained by placing the cursor over them). The lower horizontal bar represents the block found with the selected algorithm. The blocks are displayed graphically with brown rectangles going from the first to the last SNP within the block. When the cursor is over the rectangles, a tooltip text pops up in the block showing the SNPs and the haplotypes (with HapMap frequencies in parentheses). Tag SNPs are signalled with an exclamation mark (!).

DISCUSSION

It is believed that improved genotyping methods in combination with the proper bioinformatics design strategies will offer better opportunities for the study of complex diseases (3). The use of functional SNPs could be an important factor in increasing the sensitivity of association tests. Different bioinformatics approaches have been focused mainly on the effect of coding SNPs, but also recently on SNPs affecting the regulation or the splicing of genes (14).

PupasView is the first tool that integrates both transcriptional and translational phenotypic effects caused by polymorphisms. It provides an interactive environment in which functional information and population frequency data can be used over LD parameters as sequential filters to obtain a final list of SNPs optimal for genotyping purposes.

PupasView is closely linked to our previous program PupaSNP (14), which is a tool for selecting SNPs with putative

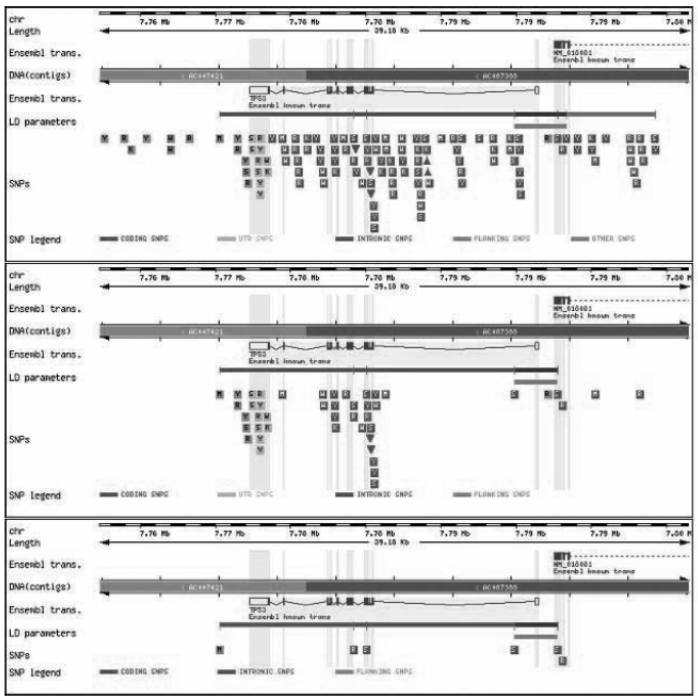

Figure 1. Sequential application of filters in PupasView. (A) SNPs in gene TP53. (B) SNPs together with population frequencies. (C) SNPs with any functional characteristic. Depending on the versions of Ensembl and dbSNP, the appearance of the figure can change.

phenotypic effects. PupaSNP, designed for high-throughput experiments, has been used to design >9000 sets of SNPs, and has a daily average of 50 uses. PupasView assists in the last refinement step of gene-by-gene selection of SNPs. Figure 1 illustrates the effect of applying successive filter steps, which are, conceptually, first to select only those SNPs which are real (with reported population frequencies) and then to select only functional SNPs. In the last view (Figure 1C), LD parameters can be used to help in the final selection.

More than 5000 SNPs have been selected using PupaSNP and PupasView in the first step of the pipeline for the study of polymorphisms at the Spanish National Genotyping Centre (CeGen).

ACKNOWLEDGEMENTS

L.C. and this work are supported by grant PI020919 from the FIS. J.M.V. is supported by the FPU fellowship programme from the MEC. This work is also partly supported by a grant from the Fundació La Caixa and the Fundación Ramón Areces. The Functional Genomics and Structure and Modelling nodes of the INB are funded by the Fundación Genoma España. CeGen, also funded by the Fundación Genoma España, is currently using the PupaSNP and PupasView programs for high-throughput SNP selection. Funding to pay the Open Access publication charges for this article was provided by xxxxx.

Conflict of interest statement. None declared.

REFERENCES

1. Collins,F.S., Green,E.D., Guttmacher,A.E. and Guyer,M.S. (2003) A vision for the future of genomics research. *Nature*, **422**, 835–847.
2. Risch,N.J. (2000) Searching for genetic determinants in the new millennium. *Nature*, **405**, 847–856.
3. Botstein,D. and Risch,N. (2003) Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nat. Genet.*, **33**, 228–237.
4. Badano,J.L. and Katsanis,N. (2002) Beyond Mendel: an evolving view of human genetic disease transmission. *Nat. Rev. Genet.*, **3**, 779–789.
5. Strittmatter,W.J., Saunders,A.M., Schmechel,D., Pericak-Vance,M., Enghild,J., Salvesen,G.S. and Roses,A.D. (1993) Apolipoprotein E: high-vidity binding to beta-amyloid and increased frequency of type 4 allele in late-onset familial Alzheimer disease. *Proc. Natl Acad. Sci. USA*, **90**, 1977–1981.
6. Hugot,J.P., Chamaillard,M., Zouali,H., Lesage,S., Cezard,J.P., Belaiche,J., Almer,S., Tysk,C., O'Morain,C.A., Gassull,M. *et al.* (2001) Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. *Nature*, **411**, 599–603.
7. Sunyaev,S., Ramensky,V., Koch,I., Lathe,W., Kondrashov,A.S. and Bork,P. (2000) Prediction of deleterious human alleles. *Hum. Mol. Genet.*, **10**, 591–597.
8. Ng,P.C. and Henikoff,S. (2001) Predicting deleterious amino acid substitutions. *Genome Res.*, **11**, 863–874.

9. Miller,M.P. and Kumar,S. (2001) Understanding human disease mutations through the use of interspecific genetic variation. *Hum. Mol. Genet.*, **10**, 2319–2328.
10. Chasman,D. and Adams,R.M. (2001) Predicting functional consequences of non-synonymous single nucleotide polymorphisms: structure-based assessment of amino acid variation. *J. Mol. Biol.*, **307**, 683–706.
11. Guerois,R., Nielsen,J.E. and Serrano,L. (2002) Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. *J. Mol. Biol.*, **320**, 369–387.
12. Ferrer-Costa,C., Orozco,M. and de la Cruz,X. (2002) Characterization of disease-associated single amino acid polymorphisms in terms of sequence and structure properties. *J. Mol. Biol.*, **315**, 771–786.
13. Ferrer-Costa,C., Orozco,M. and de la Cruz,X. (2004) Sequence-based prediction of pathological mutations. *Proteins*, **57**, 811–819.
14. Conde,L., Vaquerizas,J.M., Santoyo,J., Al-Shahrour,F., Ruiz-Llorente,S., Robledo,M. and Dopazo,J. (2004) PupaSNP Finder: a web tool for finding SNPs with putative effect at transcriptional level. *Nucleic Acids Res.*, **32**, W242–W248.
15. Hudson,T.J. (2003) Wanted: regulatory SNPs. *Nat. Genet.*, **33**, 439–440.
16. Yan,H., Yuan,W., Velculescu,V.E., Vogelstein,B. and Kinzler,K.W. (2002) Allelic variation in human gene expression. *Science*, **297**, 1143.
17. Prokunina,L., Castillejo-Lopez,C., Oberg,F., Gunnarsson,I., Berg,L., Magnusson,V., Brookes,A.J., Tentler,D., Kristjansdottir,H., Grondal,G. *et al.* (2002) A regulatory polymorphism in *PDCD1* is associated with susceptibility to systemic lupus erythematosus in humans. *Nat. Genet.*, **32**, 666–669.
18. Hoogendoorn,B., Coleman,S.L., Guy,C.A., Smith,K., Bowen,T., Buckland,P.R. and O'Donovan,M.C. (2003) Functional analysis of human promoter polymorphisms. *Hum. Mol. Genet.*, **12**, 2249–2254.
19. Colapietro,P., Gervasini,C., Natacci,F., Rossi,L., Riva,P. and Larizza,L. (2003) NF1 exon 7 skipping and sequence alterations in exonic splice enhancers (ESEs) in a neurofibromatosis 1 patient. *Hum. Genet.*, **113**, 551–554.
20. 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.
21. Krawczak,M., Reiss,J. and Cooper,D.N. (1992) The mutational spectrum of single base-pair substitutions in mRNA splice junctions of human genes: causes and consequences. *Hum. Genet.*, **90**, 41–54.
22. Kel,A.E., Gössling,E., Reuter,I., Cheremushkin,E., Kel-Margoulis,O.V. and Wingender,E. (2003) MATCH: a tool for searching transcription factor binding sites in DNA sequences. *Nucleic Acids Res.*, **31**, 3576–3579.
23. Wingender,E., Chen,X., Hehl,R., Karas,H., Liebich,I., Matys,V., Meinhardt,T., Prüss,M., Reuter,I. and Schacherer,F. (2000) TRANSFAC: an integrated system for gene expression regulation. *Nucleic Acids Res.*, **28**, 316–319.
24. Stabenau,A., McVicker,G., Melsopp,C., Proctor,G., Clamp,M. and Birney,E. (2004) The Ensembl core software libraries. *Genome Res.*, **14**, 929–933.
25. Goni,J.R., de la Cruz,X. and Orozco,M. (2004) Triplex-forming oligonucleotide target sequences in the human genome. *Nucleic Acids Res.*, **32**, 354–360.
26. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
27. Schwartz,S., Kent,W.J., Smit,A., Zhang,Z., Baertsch,R., Hardison,R.C., Haussler,D. and Miller,W. (2003) Human–mouse alignments with BLASTZ [Erratum (2004) *Genome Res.*, **14**, 786.]. *Genome Res.*, **13**, 103–107.
28. Barrett,J.C., Fry,B., Maller,J. and Daly,M.J. (2005) Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics*, **21**, 263–265.
29. Gabriel,S.B., Schaffner,S.F., Nguyen,H, Moore,J.M., Roy,J., Blumenstiel,B., Higgins,J., DeFelice,M., Lochner,A., Faggart,M. *et al.* (2002) The structure of haplotype blocks in the human genome. *Science*, **296**, 2225–2259.
30. Wang,N., Akey,J.M., Zhang,K., Chakraborty,R. and Jin,L. (2002) Distribution of recombination crossovers and the origin of haplotype blocks: the interplay of population history, recombination and mutation. *Am. J. Hum. Gen.*, **71**, 1227–1234.
31. The International HapMap Consortium (2003) The International HapMap Project. *Nature*, **426**, 789–796.

V. BIBLIOGRAFIA DEL CAPÍTOL

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 25, 3389-402.

Bateman, A., Birney, E., Cerruti, L., Durbin, R., Etwiller, L., Eddy, S. R., Griffiths-Jones, S., Howe, K. L., Marshall, M. & Sonnhammer, E. L. (2002). The Pfam protein families database. *Nucleic Acids Res* 30, 276-80.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. (2000). The Protein Data Bank. *Nucleic Acids Res* 28, 235-42.

Cargill, M., Altshuler, D., Ireland, J., Sklar, P., Ardlie, K., Patil, N., Shaw, N., Lane, C. R., Lim, E. P., Kalyanaraman, N., Nemesh, J., Ziaugra, L., Friedland, L., Rolfe, A., Warrington, J., Lipshutz, R., Daley, G. Q. & Lander, E. S. (1999). Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nat Genet* 22, 231-8.

Chakravarti, A. (2001). To a future of genetic medicine. *Nature* 409, 822-3.

Chasman, D. & Adams, R. M. (2001). Predicting the functional consequences of non-synonymous single nucleotide polymorphisms: structure-based assessment of amino acid variation. *J Mol Biol* 307, 683-706.

Collins, F. S., Brooks, L. D. & Chakravarti, A. (1998). A DNA polymorphism discovery resource for research on human genetic variation. *Genome Res* 8, 1229-31.

Ferrer-Costa, C., Orozco, M. & de la Cruz, X. (2002). Characterization of Disease-associated Single Amino Acid Polymorphisms in Terms of Sequence and Structure Properties. *J Mol Biol* 315, 771-86.

Ferrer-Costa, C., Orozco, M. & De La Cruz, X. (2004). Sequence-based prediction of pathological mutations. *Proteins* 57, 811-819.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J. P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J. C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S.,

Waterston, R. H., Wilson, R. K., Hillier, L. W., McPherson, J. D., Marra, M. A., Mardis, E. R., Fulton, L. A., Chinwalla, A. T., Pepin, K. H., Gish, W. R., Chissoe, S. L., Wendl, M. C., Delehaunty, K. D., Miner, T. L., Delehaunty, A., Kramer, J. B., Cook, L. L., Fulton, R. S., Johnson, D. L., Minx, P. J., Clifton, S. W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J. F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., et al. (2001). Initial sequencing and analysis of the human genome. *Nature* 409, 860-921.

Ng, P. C. & Henikoff, S. (2002). Accounting for human polymorphisms predicted to affect protein function. *Genome Res* 12, 436-46.

Rost, B. & Sander, C. (1993). Prediction of protein secondary structure at better than 70% accuracy. *J Mol Biol* 232, 584-99.

Sachidanandam, R., Weissman, D., Schmidt, S. C., Kakol, J. M., Stein, L. D., Marth, G., Sherry, S., Mullikin, J. C., Mortimore, B. J., Willey, D. L., Hunt, S. E., Cole, C. G., Coggill, P. C., Rice, C. M., Ning, Z., Rogers, J., Bentley, D. R., Kwok, P. Y., Mardis, E. R., Yeh, R. T., Schultz, B., Cook, L., Davenport, R., Dante, M., Fulton, L., Hillier, L., Waterston, R. H., McPherson, J. D., Gilman, B., Schaffner, S., Van Etten, W. J., Reich, D., Higgins, J., Daly, M. J., Blumenstiel, B., Baldwin, J., Stange-Thomann, N., Zody, M. C., Linton, L., Lander, E. S. & Altshuler, D. (2001). A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* 409, 928-33.

Saunders, C. T. & Baker, D. (2002). Evaluation of structural and evolutionary contributions to deleterious mutation prediction. *J Mol Biol* 322, 891-901.

Sayle, R. A. & Milner-White, E. J. (1995). RASMOL: biomolecular graphics for all. *Trends Biochem Sci* 20, 374.

Sunyaev, S., Ramensky, V., Koch, I., Lathe, W., 3rd, Kondrashov, A. S. & Bork, P. (2001). Prediction of deleterious human alleles. *Hum Mol Genet* 10, 591-7.

Sunyaev, S. R., Lathe, W. C., 3rd, Ramensky, V. E. & Bork, P. (2000). SNP frequencies in human genes an excess of rare alleles and differing modes of selection. *Trends Genet* 16, 335-7.

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., Gocayne, J. D., Amanatides, P., Ballew, R. M., Huson, D. H., Wortman, J. R., Zhang, Q., Kodira, C. D., Zheng, X. H., Chen, L., Skupski, M., Subramanian, G., Thomas, P. D., Zhang, J., Gabor Miklos, G. L., Nelson, C., Broder, S., Clark, A. G., Nadeau, J., McKusick, V. A., Zinder, N., Levine, A. J., Roberts, R. J., Simon, M., Slayman, C., Hunkapiller, M., Bolanos, R., Delcher, A., Dew, I., Fasulo, D., Flanigan, M., Florea, L., Halpern, A., Hannenhalli, S., Kravitz, S., Levy, S., Mobarry, C., Reinert, K., Remington, K., Abu-Threideh, J., Beasley, E., Biddick, K., Bonazzi, V., Brandon, R., Cargill, M., Chandramouliswaran, I., Charlab, R., Chaturvedi, K., Deng, Z., Di Francesco, V., Dunn, P., Eilbeck, K., Evangelista, C., Gabrielian, A. E., Gan, W., Ge, W., Gong, F., Gu, Z., Guan, P., Heiman, T. J., Higgins,

M. E., Ji, R. R., Ke, Z., Ketchum, K. A., Lai, Z., Lei, Y., Li, Z., Li, J., Liang, Y., Lin, X., Lu, F., Merkulov, G. V., Milshina, N., Moore, H. M., Naik, A. K., Narayan, V. A., Neelam, B., Nusskern, D., Rusch, D. B., Salzberg, S., Shao, W., Shue, B., Sun, J., Wang, Z., Wang, A., Wang, X., Wang, J., Wei, M., Wides, R., Xiao, C., Yan, C., et al. (2001). The sequence of the human genome. *Science* 291, 1304-51.

Wang, Z. & Moult, J. (2003). Three-dimensional structural location and molecular functional effects of missense SNPs in the T cell receptor Vbeta domain. *Proteins* 53, 748-57.