Capítol 6.

Ús d'eines bioinformàtiques per l'anotació de mutacions puntuals en models animals.

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

I. INTRODUCCIÓ

A continuació de la seqüenciació del genoma humà han seguit la de varis organismes model com Drosophila (Adams et al., 2000), ratolí (Waterston et al., 2002), etc. que han tingut com a conseqüència l'obtenció d'una gran quantitat de dades, entre elles, els SNPs, tant d'humans com d'altres espècies. L'innegable impacte que aquesta informació pot tenir sobre els estudis de biomedicina ha constituït un afegit pel desenvolupament d'eines d'anotació d'aquesta variabilitat.

D'entre els organismes model seqüenciats destaca el ratolí que s'ha usat durant molt de temps com a model en l'estudi de les malalties humanes generant-se durant tot aquest temps moltes dades d'origen bioquímic i genètic. Per tant la seqüenciació del genoma del ratolí i l'estudi de la seva variabilitat ha esdevingut un eina d'interès en l'estudi de les malalties humanes. Aquest interès s'ha vist incrementat per la similaritat entre els genomes de les dues espècies: la identitat de seqüència entre proteïnes homòlogues de ratolí i humà està entre un 80% i 100% (Waterston et al., 2002), el que implica un grau de conservació funcional i estructural molt important (Russell & Barton, 1994). En aquest context, l'estudi de la variabilitat en ratolí pot ajudar a esclarir la relació de la variabilitat del genoma humà i la malaltia (Rossant & McKerlie, 2001). Aquesta possibilitat ha donat lloc al desenvolupament d'estudis a gran escala de mutagènesi sobre ratolí per tal d'identificar gens relacionats amb patologies (Hrabe de Angelis et al., 2000; Nolan et al., 2000) amb la idea de transferir la informació obtinguda a les patologies humanes. Aquests model de treball també és aplicable a altres animals models com la rata (Gibbs et al., 2004) o la mosca del vinagre (Adams et al., 2000; Celniker et al., 2002). En aquest context han pres un valor especial les eines bioinformàtiques per la caracterització de les mutacions, degut al seu creixent precisió i al seu cost, pràcticament nul.

Actualment s’han desenvolupat diferents aproximacions per l’anotació de variacions puntuals (Chasman & Adams, 2001; Conde et al., 2004; Ferrer-Costa et al., 2004; Ng & Henikoff, 2001; Santibañez-Koref et al., 2003; Sunyaev et al., 2001; Wang & Moult, 2001). Algunes incideixen en mapar sobre la seqüència de DNA determinant especialment quin tipus de regions són. Altres es centren en SNPs que afecten regions codificants on intenten predir l’efecte de les mutacions en la funció, relacionant-lo amb l’efecte patològic.

Tanmateix la transferència dels mètodes, esmentats en capítols anteriors, a altres espècies no és simple. Alguns treballs mostren que l’ús de dades provinents d’altres espècies, o amb molt poques proteïnes, poden donar resultats esbiaixats (Ferrer-Costa et al., 2004) que redueixen el rang d’aplicació d’aquests mètodes. Un altre factor que dificulta la transferibilitat és la divergència de les seqüències en l’especiació, així aquesta divergència pot acollir mutacions sense problemes en una espècia mentre que en altres són patològiques (Gibbs et al., 2004; Huang et al., 2004; Kondrashov et al., 2002; Waterston et al., 2002). Aquest fenomen es pot explicar per l’aparició de mutacions compensatòries, fent que els mètodes de predicció puguin fallar.

En aquest treball s’estudia l’aplicabilitat a altres espècies d’un mètode d’anotació de DAMUs basat en mutacions humanes. S’estudia el grau de similaritat entre seqüències ortòlogues requerit per l’aplicació del mètode sense la pèrdua de capacitat predictiva. Addicionalment, i dins d’aquest context, es comparen mutacions humanes i no humanes en relació a les propietats usades per puntuar-les en diferents mètodes de predicció.

II. MATERIALS I MÈTODES

Per veure quan mutacions no humanes poden ser anotades usant eines entrenades amb mutacions humanes es va usar el mètode de predicció (Ferrer-Costa et al., 2004) desenvolupat al nostre laboratori i explicat en capítols anteriors. Es resumeix a continuació les principals característiques del mètode predictiu, tanmateix es troba més informació en dos capítols anteriors el tercer (Ferrer-Costa et al., 2002) i el quart (Ferrer-Costa et al., 2004).

Descriptors de les mutacions

Els paràmetres usats en aquest treball es van obtenir sempre usant només informació de seqüència i es poden classificar en tres grans categories: i) descriptors relacionats amb l'estructura, ii) propietats de residu/seqüència, i iii) propietats evolutives.

Tots ells han estat descrits amb deteniment en els capítols 3 i 4.

Els grups de dades de mutacions

Seguint treballs anteriors (Ferrer-Costa et al., 2002; Ferrer-Costa et al., 2004; Goodstadt & Ponting, 2001; Ng & Henikoff, 2001; Sunyaev et al., 2001), les mutacions patològiques (DAMUs) es van obtenir de la base de dades SwissProt (Apweiler et al., 2004). Es va usar la versió 40 i es van obtenir les mutacions preguntant a la base de dades per les paraules claus: DISEASE, VARIANT i HUMAN/MOUSE. En etapes posteriors de filtració, les mutacions que no tenien un vincle clar amb malaltia es van rebutjar, també es van eliminar aquelles mutacions per les quals no hi havia alineament en la base de dades Pfam (Bateman et al., 2002). Aquest procediment va rendir 9334 mutacions patològiques sobre 811 proteïnes humanes. La quantitat de dades per ratolí és menor, amb 75 mutacions patològiques mapades en 54 proteïnes. Aquest procediment es va repetir generant un grup de DAMUs no humanes, en les quals es va afegir a les mutacions de ratolí nou espècies més: vaca, gos, cavall, porc, rata, conill, xai, visó i guilla. El nombre total de mutacions per aquest segon grup de dades es de 105 DAMUs sobre 79 proteïnes.

Per tal d'explorar fins a quin punt el grup de dades derivat de proteïnes de ratolí era representatiu de l'espai funcional cobert per les proteïnes conegudes de ratolí, es van comparar la llista de proteïnes de ratolí amb totes les proteïnes de ratolí presents a SwissProt (fins un total de 8337 proteïnes). Amb aquesta idea, es van anotar les dues llistes de proteïnes amb termes de GO (*Gene Ontology Database*) (Harris et al., 2004) usant el programari FatiGO (Al-Shahrour et al., 2004), que també dóna comparació estadística entre les dues llistes tenint en compte el problema de tempteig múltiple. Fent la comparació usant el terme *Biological Process* en nivell dos, es pot veure que malgrat les diferències de mida dels grups comparats, el grup de proteïnes de ratolí que contenen mutacions patològiques reprodueix de manera raonable les principals característiques del grup amb totes les proteïnes de ratolí. La mateixa comparativa es va fer pel nivell tres de GO per tal d'obtenir una classificació més precisa. Els resultats són similars malgrat que com és obvi el mapatge del grup de proteïnes amb mutacions patològiques sobre els termes GO és més dispers.

Per les mutacions neutres (*NEMUs*) es va seguir el model evolutiu estàndard (Ferrer-Costa et al., 2002; Ferrer-Costa et al., 2004; Santibañez-Koref et al., 2003; Sunyaev et al., 2001), que defineix una mutació com a neutre quan apareix en organismes molt propers evolutivament. Es van obtenir les *NEMUs* humanes i de manera anàloga les de ratolí, de la següent manera. Es van prendre els alineament múltiples Pfam (Bateman et al., 2002) per la proteïna humana i es van eliminar: i) totes les altres seqüències humanes; (ii) totes les seqüències no humanes amb una identitat de seqüència inferior del 95% a la proteïna estudiada. Qualsevol canvi aminoacídic entre la seqüència humana i les altres seqüències es va considerar com a mutació neutra (*NEMU*). El mateix procediment es va aplicar a ratolí i les altres espècies. Al final es van eliminar les mutacions neutres que eren comunes als dos grups de *NEMUs*. Al final vam obtenir 11374 mutacions neutres humanes, 373 mutacions neutres de ratolí i 888 *NEMUs* pel grup de proteïnes no-humanes.

Les xarxes neurals

Basat en treballs anteriors mostrats en els capítols anteriors (Ferrer-Costa et al., 2004) es van usar una xarxa neural del tipus *feed-forward* (NN) (Rumelhart et al., 1986) amb una capa d'entrada, una capa de sortida i una capa oculta amb dues unitats. Aquest model ens permet mantenir la relació entre les mutacions d'entrenament i els pesos de la xarxa entre 5 i 10 per evitar sobreentrenament quan es fan servir conjunts de dades petits (Mehrotra et al., 1997).

Dos grups de paràmetres es van usar com a entrada, un incloïa els 15 paràmetres descrits anteriorment, el segon incloïa només dos variables: PSSM i les mesures d'energia lliure de transferència entre aigua i octanol (Fauchere & Pliska, 1983). Aquests dos paràmetres són relativament independents i han mostrat un poder discriminatori significant (Ferrer-Costa et al., 2004) i un rendiment òptim quan es compara amb altres combinacions de dos paràmetres. En alguns casos, i com a referència, es van usar tres paràmetres com a entrada, PSSM (eq. 4), MMS (eq. 3) i valors de PAM40 que mostren un gran poder informatiu quan s'usen junts. Els pesos de la xarxa es van optimitzar usant gradients conjugats escalats en 500 iteracions. El programari de les xarxes neurals usat es FFNN, amablement cedit pel Dr. Adrian Shepherd de la University College of London.

Per una mutació donada, la sortida de la xarxa és un número comprès entre 0 i 1. La mutació es predita com a patològica per valors per sobre 0.5 i neutres per valors per sota.

Mesures de rendiment

Les mesures del rendiment del procediment de predicció es va mesurar usant 4 índexs diferents. Tots ells detalladament descrits en el capítol 4.

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

III. ARTICLE DE RECERCA

Use of bioinformatics tools for the annotation of disease-associated mutations in animal models. Carles Ferrer-Costa, Modesto Orozco, Xavier de la Cruz. Proteins, doi: Not yet available, accepted for publication June 7 2005.

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

USE OF BIOINFORMATICS TOOLS FOR THE ANNOTATION OF DISEASE-ASSOCIATED MUTATIONS IN ANIMAL MODELS

Carles Ferrer-Costa[1], Modesto Orozco [1,2*], Xavier de la Cruz [1,3*]

[1] Molecular Modeling and Bioinformatics Unit. Institut de Recerca Biomédica. Parc Científic de Barcelona. Josep Samitier 1-5. 08028 Barcelona, Spain.

[2] Departament de Bioquímica i Biologia Molecular. Facultat de Química. Universitat de Barcelona. Martí i Franquees 1. 08028 Barcelona, Spain.

[3]Institució Catalana per la Recerca i Estudis Avançats (ICREA), Passeig Lluís Companys, 23, 08018 Barcelona, Spain.

* **Send correspondence** to Xavier de la Cruz or Modesto Orozco. Parc Científic de Barcelona; C/Josep Samitier, 1-5; 08028 Barcelona; SPAIN. Fax: +34 93 403 71 57. Email: xavier@mmb.pcb.ub.es/modesto@mmb.pcb.ub.es.

Running Title : Prediction of human and mouse pathological SNPs

KEYWORDS: animal models, disease, disease-associated mutations, bioinformatics, sequence variability

ABSTRACT

Single-point mutations are one of the most frequent causes of genetic variability in both human and close species. The recent availability of different bioinformatics tools for annotating human SNPs has opened the possibility of using them to score SNPs from species with a biomedical interest, in particular from mouse and other models of human disease. Also, this ability to predict pathogenicity of single point mutations in one species, based on data from another species, opens the possibility to predict the pathological character of single point mutations in human using data from well characterised model systems of human disease. This could provide a valuable alternative to the more traditional genetic population approaches. However, transferral of prediction tools may be limited by different factors, from a species biass in the training set, to a large sequence divergence between the proteomes of the training and the target species.

Here we study the conditions under which prediction tools can be transferred among species, concentrating in the case of mouse. We find that for the majority of the human-mouse homolog pairs the sequence similarity is large enough to preserve the pathological character of mutations among species, in general. We then establish that prediction/annotation tools developed for one organism can be used to predict the neutral/pathological character of mutations/SNPs in the other organism.

INTRODUCTION

Single nucleotide polymorphisms (SNP) constitute the main source of genetic variation in humans [1]. While many of them do not have an impact in the organism's health, others can lead to disease, either when alone (monogenic diseases) or in combination with mutations in other genes (polygenic diseases) [1,2]. The sequencing of the human genome [3,4], together with the existence of different genotyping projects [5,6] have generated a massive amount of experimental data on human SNPs, thus increasing the need for annotation tools. Unfortunately, experimental characterization of SNPs as neutral or pathological is much more difficult than finding them, even in cases of monogenic diseases [7,8]. This is becoming a serious problem as the gap between amount of information and its applicability for the study of human disease grows [9].

In these circumstances, the sequencing of the mouse genome and the study of its variability has become an important advance [9]. Mice have constituted valuable tools for the study of human disease for many years [10], and a large amount of biochemical, genetic and disease-related information is already available for them. In addition, and apart from its obvious technical advantages, the use of murine models is also supported by the close similarity between the genomes of both species [11]. For example, at the protein level, sequence identity between human sequences and their mouse orthologues mainly varies between 80 % and 100 % (Figure 1). This degree of similarity implies a high degree of structural and functional conservation for the involved proteins [12]. Within this context, the study of their sequence variability and its relation to disease may help to bridge the gap between human genome information and human disease [9]. This interest has also triggered the development of large-scale ENU mutagenesis studies [13,14], where a large number of mutant mice can be generated. These mice are then subject to extensive phenotypic screens, to try to identify diseased subjects and relate their pathology to the individual mutation. This information could subsequently be utilised to infer the effect of these mutations in human, provided that quantitative techniques exist to transfer mutation information between species.

In addition to mouse, genome information is also available for other animal models, like rat [15] or fruitfly [16,17]. In the case of rat, an also valuable model of human disease, the annotation of sequence variability may bring large benefits in the identification of disease-associated, or disease-causing, variants. Within this context it is worth mentioning the pioneering work by Cuppen and coworkers who have utilised two bioinformatics tools [18], Polyphen [19] and SIFT [20] developed for the prediction of human DAMUs, as an alternative to experimental annotation of SNPs.

At present, there are several bioinformatics approaches for the annotation of human DAMUs [7,8,20-24]. In some cases, they provide a mapping of mutations/SNPs along the DNA, together with the functional type of the affected region, e.g. regulatory, coding, etc [22]. Other strategies focus in non-synonymous mutations happening in protein coding regions, thus working with a necessarily small number of cases, but providing the user with a more precise information on their pathological character [7,8,20,21,23,24]. The latter methods are based on the use of different tools, like neural networks, trained in datasets of both neutral and disease-associated mutations, and able to identify the possible pathological character of a target mutation/SNP utilising a series of sequence or structure-derived properties. Although the success rate of these methods vary, it is usually above 70 %, e.g. the approach developed in our group [25] reaches 84% cross-validated success rate with human DAMUs, using only sequence-derived information, with a 66.5 % improvement over random predictions. Overall, these programs constitute a simple and cheap approach to the prediction of DAMUs, and they are being used for this purpose by different authors [26,27].

However, transferring these prediction methods and their results from human to other species, or *vice versa*, is not straightforward. First, most have been trained utilising data on human DAMUs, or a very small number of proteins, which may bias the resulting parameters and lead to a poorer performance when applied to other species. This problem has been discussed by Westhead and coworkers [28] for the annotation of DAMUs using artificial intelligence tools; and also by de la Cruz and Calvo [29], within the context of the prediction of protein stability changes associated to single-point mutations. The problem may become worse if we focus in specific systems of biomedical interest, like p53 or hemoglobin. In this case, the large number of disease-associated mutations and the relevance of the system justify the development of specific tools[25,30]. However, the improvement in prediction performance comes at the expense of biasing the method towards the properties of the protein studied, including a species biass. For example, if we derive our method using p53 data[25] and use it to score hemoglobin mutations, we find a 12 % decrease in the performance, relative to the cross-validated performance of the same method when trained on hemoglobin mutations[25]. These data highlight the existence of training biasses that may limit the transferral of prediction methods either between systems or among species.

A second factor that can limit the transferability of prediction tools among species is the fact that sequence divergence may lead to the appearance as wild-type residues in one species of amino acids that would be pathological in other species, at an equivalent sequence position [31]. This possibility was mentioned in an earlier article by Pauling and Zuckerkandl [32] who describe the presence in the orang-utan haemoglobin of an amino acid that is associated to the Norfolk syndrome when present in the human haemoglobin. Recently, the sequencing of the mouse and rat genomes has revealed more cases where this situation happens [11,15,33,34]. For example, the Val(198)→Met in cryopyrin is associated to the familial cold autoinflammatory syndrome in humans, while Met is found at an equivalent position in both mouse and rat wild-type sequences [34]. More general studies indicate that this may be a relatively common situation [33] and that mutations that are pathological in one species might not be in others. This behaviour is probably due to the presence of compensatory mutations in the protein [33,35]. At a practical level, this effect may eventually result in the failure of prediction methods, e.g. neutral alleles scored as pathological and *vice versa*, in particular when the target protein has diverged substantially from other family members.

In this article we study the applicability to other species, in particular to model animals like mouse, of human-based methods for the annotation/prediction of DAMUs. We first study under which conditions this can be done, that is, at which point sequence divergence may affect the power of prediction methods. Subsequently, we compare human and non-human mutations in relation to the properties normally utilised to score them by the different prediction methods. Then, using our own neural network-based method [25] we quantify how well we can identify non-human DAMUs. Our results indicate that a minimum sequence identity of 70 % - 80 % guarantees a transferral of the prediction tools among species with no significant changes in their prediction performance. This is confirmed by the fact that the interspecies prediction works reasonably well, with an overall success rate around 85 %, and an improvement over random between 45 (set of diverse non-human mutations) to 50 % (set of only mouse mutations). Interestingly, the reverse is also true, and non-human trained networks predict remarkably well human DAMUs. This finding opens the possibility to use animal models, where rare mutations can be engineered, to predict the impact of single point mutations in human health.

METHODS

To see whether non-human mutations can be annotated using human-trained tools, we have utilised our own prediction method [25], that shares some characteristics with well known methods like Polyphen [19] and SIFT [20], but with a higher cross-validated average performance[25]. Next, we summarize the main characteristics of our predictive model, an explanation that can also be found elsewhere with slightly more detail [7,25].

MUTATION DESCRIPTORS

The parameters utilised in this paper were always obtained using only sequence information and can be classified in three main categories: i) structure-related descriptors, ii) residue/sequence properties, and iii) properties derived from multiple sequence alignments.

Structural-related descriptors. Predictions of the secondary structure and solvent accessibility at the mutation site, were obtained from the protein sequence using the PHD software package [36] (www.embl-heidelberg.de/predictprotein) with default parameters. The three secondary structure states, helix, beta and coil, were encoded as 0, 1 and 2, respectively. It has to be mentioned that encoding secondary structure using three-element vectors, e.g. helix (1, 0, 0), beta (0, 1, 0) and coil (0, 0, 1), constitutes a better practice. However, we discarded this option because of the need to keep a low number of parameters, and because our previous data showed that the discriminant power of secondary structure is very low [7].

The three accessibility states, buried, half-buried and exposed, were encoded as 0, 1 and 2, respectively.

Residue/Sequence properties. The physico-chemical changes induced by the mutation are measured using three types of parameters that correspond to the following properties: i) hydrophobicity, ii) secondary structure propensity, and iii) volume. Changes in hydrophobicity were described utilising two scales: water/octanol free energy measurements [37], and statistical potentials [38]. Two scales were utilised for secondary structure propensities [39,40]. Size changes were described using van der Waals volumes [41] and volume of buried residues [42].

Blosum62 [43] and PAM40 [44] mutation matrices were also used to label mutations, as they include different contributions that may be related to the damage caused by a mutation. It has to be mentioned that utilising the raw Blosum62 matrix may affect the estimates of variability. However, the impact of this effect in the method's performance is implicitly taken into account in the testing procedure followed.

We also utilised a simple sequence potential, Ptseq(r_j) (see eq. 1), related to the probability of observing residue r_j at position j, in a given sequence environment:

$$\mathit{Ptseq(r_j)} = \ln\left[\prod_{i=-5}^{5} \mathrm{n(r_j, r_{j+i})/n(r_{j+i})}\right] \qquad (1)$$

where $n(r_j,r_{j+i})$ is the number of pairs of amino acids of types r_j and r_{j+i} at a sequence distance i. $n(r_{j+i})$ is the total number of residues of type r_{j+i}. These numbers are computed using the whole set of human sequences from the SwissProt database [45].

Properties derived from multiple sequence alignments. We utilised two measures of the amino acid variability, and two position-specific scores, all derived from Pfam [46] multiple sequence alignments.

Variability at the mutation site was determined using Shannon entropy [47] (see eq. 2) and an average of mutation matrix scores [30] (see eq. 3).

$$ShS = \sum_i P_{ij} \ln_2 P_{ij} \tag{2}$$

where P_{ij} is determined as the frequency of amino acid i at position j. The sum runs over the 20 amino acids defining the mutation space.

$$\left[\frac{\sum_{k=1}^{N} \sum_{z=k+1}^{N} s_{kz}}{\frac{N!}{2 \times (N-2)!}} \right] \Big/ \mathrm{S}_{\mathrm{max}} \tag{3}$$

where s_{kz} is the element of the raw Blosum62 matrix [43] corresponding to the comparison between residues in sequences k and z in the multiple sequence alignment, at the position of the mutation under study. S_{max} is the larger s_{kz}, and N is the number of sequences in the alignment.

Two types of position-specific scoring parameters, based on modified versions of the log-odds ratio [25,48] were used (eqs 4 and 5). Both indexes emphasize the importance of multiple alignment information when it includes a large number, N, of sequences, and converge to Blosum matrices otherwise.

$$\frac{N\sigma}{1+N\sigma} \log(\mathrm{p_{mj}}/\mathrm{p_m}) + \frac{1}{1+N\sigma} \mathrm{B62_{wm}} \tag{4}$$

where N is the number of sequences in the multiple sequence alignment, σ is taken as 0.02 [48], w and m stand for the normal and mutant amino acid, respectively. $B62_{wm}$ is the element of the Blosum62 matrix [43] corresponding to the mutation from the wild-type residue (w) to the mutant residue (m).

$$\frac{N\sigma}{1+N\sigma}\left[\log(p_{mj}/p_m) - \log(p_{wj}/p_w)\right] + \frac{1}{1+N\sigma}B62_{wm} \qquad (5)$$

where m, w, N, σ, p_{mj}, p_m and $B62_{wm}$ have the same meaning as before. p_{wj} is the relative frequency of the original amino acid type w at position j in the multiple sequence alignment. p_w is the frequency of the same amino acid type in human sequences in SwissProt [45].

THE MUTATIONS DATA SETS

Following previous work, both from our laboratory and from other groups [7,8,20,25,49] DAMUs were obtained from Swissprot [45]. We utilised version 40 and obtained our set of DAMUs querying SwissProt with the keywords: DISEASE, VARIANT, and HUMAN/MOUSE. In a subsequent filtering step, mutations with no clear links to disease were rejected (after looking at relevant references), as well as those for which only the human and/or mouse sequence were present in the Pfam [46] alignment for the protein family. This procedure gave 9334 pathological mutations in 811 human proteins. The amount of data for mouse is smaller, with 75 pathological mutations mapping to 54 proteins. This procedure was repeated generating a set of non-human DAMUs, in which together with the mouse mutations we also included mutations from nine other species: cattle, dog, horse, mink, opossum, pig, rat, rabbit and sheep. The total number of mutations for this second set was of 105, mapping to 79 proteins.

To explore to which point the dataset of mouse proteins was representative of the functional space covered by known mouse proteins, we compared our list of mouse proteins with the whole set of mouse proteins in SwissProt (a total of 8337 proteins). To this end, we annotated both protein lists with GO terms [50], using the FatiGo software [51], that also provides a statistical comparison between the two lists, taking into account the multiple testing problem. The results for the GO "Biological Process" terms, at level 2, are shown in Figure 2, where one can see that in spite of the size differences, our set of mouse proteins reproduces reasonably well the main features of the whole mouse set. No statistically significant differences were found for any of the GO terms (data not shown). This test was also done for a finer classification level -GO level 3- with similar results, although in this case the mapping of our mouse set among the GO terms was obviously more sparse (results available upon request).

For neutral mutations (NEMUs) we followed a standard model [7,8,23,25], which defines a mutation as neutral when it appears in closely related organisms. In order to obtain the human NEMUs (the procedure for mouse was completely analogous) we took the Pfam [46] alignment for the human protein family, and eliminated: (i) all the human sequences; (ii) all non-human sequences with less than 95 % identity to the target sequence. Any amino acid change between the target human sequence and the remaining sequences was then considered a human NEMU. As mentioned before, the same procedure was followed for the mouse NEMUs and the other non-human mutations. Finally, we discarded those mutations common to both sets of neutral mutations. This procedure yielded 11374 human NEMUs, and 373 mouse NEMUs, and 888 NEMUs for the non-human dataset. All the mutation data are available on demand from the authors.

THE NEURAL NETWORK

Based on our previous studies [25] we have used, unless otherwise stated, a feed-forward neural network (NN) [52] with one input layer, one output layer and a hidden layer with two units. This model allowed us to keep the ratio between number of training mutations and neural network weights between 5 and 10, to avoid overfitting when training the mouse-based NN [53].

Two sets of parameters were used as input, one including all the 15 parameters described in the *Methods* section, and the second, including only two parameters: PSSM and water/octanol free energy measurements [37](see above). These two parameters are relatively independent, and have been shown to provide a significant discriminatory power [25] and an optimal performance when tested against other combinations of two parameters. The network weights were optimised using scaled conjugate gradients with 500 iterations. The neural network package was FFNN, kindly lent by Dr. Adrian Shepherd, at the Birkbeck College London.

For a given mutation, the network output is a number comprised between 0 and 1. The mutation is predicted as pathological for values above 0.5, and neutral for values below.

PERFORMANCE MEASURES

The performance of our prediction method was measured utilising four different indexes. First, the percentage of correct predictions (Q_{tot}; eq. 6), that provides an overall measure of the performance of the method.

$$Q_{tot} = 100 \frac{Tp + Tn}{(Tp + Tn + Fp + Fn)} \qquad (6)$$

where Tp, Tn are the number of mutations correctly predicted as DAMUs (true positives) and NEMUs (true negatives), respectively; Fp, Fn are the number of mutations incorrectly predicted as DAMUs (false positives) and NEMUs (false negatives), respectively.

The performance of the method relative to random predictions is measured with the S_{tot} index [54] (eq. 7). This index can be adapted to provide more precise information on the performance for DAMUs or NEMUs (see eqs 8a,b)

$$S_{tot} = \frac{(Tp + Tn) - R}{t - R} \times 100 \qquad (7)$$

where $t = Tp + Tn + Fp + Fn$ and $R = [(Tp + Fp).(Tp + Fn).(Tn + Fp).(Tn + Fn)] / t$

$$S^{DAMU} = \frac{Tp - R^{DAMU}}{Tp + Fn - R^{DAMU}} \times 100 \qquad (8a)$$

$$S^{NEMU} = \frac{Tn - R^{NEMU}}{Tn + Fp - R^{NEMU}} \times 100 \qquad (8b)$$

where $R^{DAMU} = [(Tp + Fp).(Tp + Fn) / t$ and $R^{NEMU} = [(Tn + Fp).(Tn + Fn)] / t$

Some neural networks were trained to discriminate between DAMUs occurring in mouse and human. This was done to test if the NN was able to discriminate between both mutation types, in order to establish a cross-species prediction method. In this case, the quality of the predictive models was assessed using the same parameters described in eqs 6-8a,b, but with super-indexes "DAMU" and "NEMU" replaced by "mouse" and "human". This procedure was repeated in an analogous fashion with human and non-human DAMUs.

RESULTS AND DISCUSSION

While there is no doubt about the overall coincidence in the main biochemical processes among human and animal models like mouse (Rossant & McKerlie, 2001), it is also clear that there is a certain amount of sequence divergence among homolog proteins. This divergence, that can be particularly fast for some families[11], may be at the origin of the several cases in which human DAMUs correspond to wild-type alleles in other species [11,32,34,35]. For this reason, the first point to study was to determine if mouse and human DAMUs occur in a similar structural/functional background, in spite of the sequence divergence between both species. That is, we wanted to know to which extent a human pathological residue at a given location is also likely to be pathological in mouse at the equivalent location, and *vice versa*. This will give us an idea about whether the corresponding proteins are close family members, from a DAMUs point of view. To explore this point we compared two distributions of sequence identities. The first corresponds to the alignments between all human and mouse SwissProt [45] homologues. The second is the sequence identity distribution between human and non-human homologues when the latter have, at least in one position, an amino acid that would be pathological if in the human protein. The overlap between both distributions, at sequence identities above 70 % - 80 %, is really minor (Figure 1) suggesting that only in a relatively small number of cases an amino acid appearing as pathological in human will be neutral in mouse. In other words, we can reasonably expect that pathological mutations in humans will also be pathological in mouse, and *vice versa*. Therefore, the performance of human-based prediction methods should hold when applied to mouse proteins, on the average. However, because disease-related genes could have a different distribution than normal genes [55], we computed a third distribution. The latter was obtained using only a subset of 105 genes associated to disease in both human and mouse, according to SwissProt [45] records. For this distribution we found essentially the same behaviour (Figure 1) as that corresponding to the general human-mouse comparison. All these data considered together support the transferral of prediction methods obtained in one species to other species. In addition, when sequence similarities between homologues are above 70% - 80%, their prediction performance in the target species is going to be very similar to that in the training species.

Next, we studied whether DAMUs themselves are similar in both species, from the point of view of the properties that will be subsequently utilised to score them. This will guarantee that these properties are suitable to score both human and other species mutations. To test this point we compared the distribution of the different parameters used as mutation descriptors in the mouse and human sets of DAMUS (see *Methods*). In all cases we found that the distribution was the same in both species (see examples in Figure 3), indicating that there was no property showing statistically significant human (or mouse) bias, and thus also supporting the use of human-based methods for the scoring of SNPs/variants in mouse. A final test on the similarity of DAMUS distributions between species was done by training a simple NN (with no hidden layers), to distinguish between DAMUs in human and mouse. The NN was trained considering an extended (15 descriptors) and also a reduced (PSSM and a hydrophobicity index, results not shown) set of parameters to reduce over-training. None of the NNs was able to discriminate between human and mouse DAMUs, all showing a zero percent improvement over random procedure (Table 1). All these findings strongly support the use of bioinformatics tools trained with mutational data in one organism to evaluate the pathogenicity of single point mutations in other organisms.

We then characterised the actual performance of human-trained NN when utilised to predict pathological mutations in other organisms, and *vice versa*. To this end, we tested the predictive power of a NN with one hidden layer and two units (see *Methods*), trained with 9334 DAMUs and 11374 human DAMUs from 811 proteins, in the mouse dataset of 75 DAMUs and 373 NEMUs, from 54 proteins. Calculations were performed for two models (see Methods), one using only two parameters (PSSM and residue hydrophobicity) as input, and the second considering the entire set of 15 descriptors.

From our previous work [25], we know that a more complex human-trained NN (19 input descriptors, with database annotations added to the mutation labels used here, and one hidden layer with 22 nodes) has a good performance when discriminating between DAMUs and NEMUs in humans [25], with around 85% correct predictions (Q_{tot}, eq. 6) and 66.5 % improvement over random (S_{tot}, eq. 7). This predictive power decreases, although not dramatically, when the two-descriptor NN was used (Q_{tot} around 79 % and S_{tot} around 56 %) is used. The figures when utilising three input parameters (see *Methods*) are similar (Q_{tot} around 80 % and S_{tot} around 59 %).

When applying the most complete human-trained NN, that using 15 descriptors as input, to the mouse set we find reasonably good results (see Table 2): 86 % predictions are correct, with a normalised improvement over random around 53 %. The performance decrease relative to the previous data can be traced to the S^{NEMU}, 49 %, which is lower than that obtained when predicting human mutations with human-trained NNs (around 70 %). This can be essentially attributed to size variations between the human and mouse samples. This can be shown simulating the S^{NEMU} probability distribution in samples with the characteristics of our mouse sample (see Appendix): we find that the human-trained NN will give an average S^{NEMU} of 44 % (±14 %), when applied to mouse samples with the same size, DAMUs and NEMUs proportions, as our mouse sample. The S^{NEMU} value for our sample, 49 %, is close to the average value, therefore indicating that the observed "decay" can be attributed to sample properties like size, and DAMUs and NEMUs proportions. Similar, or slightly worse, figures were obtained for the NN using two input parameters (Table 2): 86 % overall success rate, and 51 % improvement over random. Overall, these results confirm that human-trained NNs can be utilised to score mouse mutations, with useful success rates and improvements over random. The small performance decrease observed, relative to the human-trained NN when tested in human mutations, is mainly due to sample properties like size and composition of DAMUs and NEMUs.

We also tested whether the method could work in the absence of any homologue to the mouse proteins in the human training set. The prediction ability of the method remained, but for a slight decrease, when human homologues of the mouse proteins were eliminated from the training set: 82 % overall success rate, and 42 % improvement over random. This indicates that scoring mutations in mouse proteins does not require training the prediction method with data from the corresponding human homologues. However, it is also clear that the presence of the latter in the training set will improve any prediction process.

To further characterise the prediction process we analysed the prediction failures, to see whether there was any trend among them. To this end we went through the values of the different parameters, however, we focused our analysis in the position-specific scores, as they are the main contributors to the method's performance [25]. In the case of missed DAMUs, we find that most of their values are positive, and slightly above the decission boundary of the NN [25]. For this reason the latter will tend to predict them as neutral. This usually happens in proteins belonging to large families, where some alignment positions may be highly heterogeneous. For example, in the case of the Microphthalmia-associated transcription factor, the mutation D329N associated to Microphthalmia-vitiligo is predicted as neutral. A visual inspection of the corresponding position in the family alignment shows a large heterogeneity, with N being highly frequent in one of the sub-families. A similar explanation applies to incorrectly predicted NEMUs, which tend to have values of the position-specific score slightly below the decission boundary of the NN, thus being predicted as DAMUs. Again this situation is most likely to happen in large families. We also compared the list of GO terms [50] for the proteins corresponding to: (i) the mutations in the test set; and (ii) the subset of the latter which were incorrectly predicted. This comparison was done with the FatiGo software [51], and no statistically significant difference were found between both lists of proteins. That is, prediction failures are evenly distributed among the different functional classes.

If we now consider the reverse problem, that of predicting human mutations with a mouse-trained NN, we see that the NN with the 15 parameters input had an overall success rate of 78 %, with an improvement over random over 55 %. We see this performance is lower than that of the human-trained NN network predicting mouse mutations. In particular, there has been a reversion in the S^{NEMU} and S^{DAMU} values, which are 69 % and 46 % for the mouse-trained NN, and 49 % and 60 % for human-trained NN. As before, these differences can be mainly explained in terms of the above mentioned sample properties. If we simulate the probability distribution of the S^{NEMU} and S^{DAMU} for the human sample, we find that the average values for these two variables are: 46 % (±5 %) and 96 % (±4 %). That is, the reversion in the S^{NEMU} and S^{DAMU} values can be explained by the properties of the human sample. However, we also see a deviation for the mouse-trained S^{NEMU} value, 60 %, which is lower than the computed average, 96 ±4 %. This is probably due to the biased procedure utilised to obtain the dataset of NEMUs (see *Methods*), and to some overtraining due to the imbalance between mouse NEMUs and DAMUs. Still, it has to be mentioned that the performance of the mouse-trained NN is remarkable, considering that it was trained with only 448 mutations (DAMUs + NEMUs) happening in a small set of 54 proteins and was utilised to score, with a substantial success rate, 20708 human mutations happening in a set of 881 proteins, covering a broad range of protein functions and structures.

The simpler mouse-trained NN -the one using two parameters as input, gave comparable results, with a 77 % overall success rate and 51 % improvement over random. Similar results were obtained after eliminating human homologues of the mouse proteins. To try to push further our transferability test of the human-trained tools, we decided to increase the mouse set with some additional mutations from other mammalian models, increasing our set of non-human DAMUs to 105 (see *Methods*). The average identity for these additional sequences was: 89 % (± 7 %). The results obtained for the human-trained NN (15 parameters) were 85 % overall success rate, with a 42 % improvement over random. This slight decrease in the performance over random can be again traced to the NEMUs, and as before it can be explained in terms of the characteristics of the non-human set utilised (data not shown), although some problems in the selection procedure for the NEMUs can also have contributed. Slightly better performance is found with the two-parameter NN (Table 2): 89 % overall success rate, and 51 % improvement over random. The difference is attributable to a certain degree of overtraining in the 15-parameter NN, although the differences between both NNs are not too large.

Again, a 15-parameter NN trained excluding homologues from the training set showed a small performance decrease, with 80 % overall success rate and 34 % improvement over random. A similar trend is observed for the 2-parameter NN, with 83 % overall success rate, and 41 % improvement over random. Analysis of the individual errors gives similar results as for the prediction of mouse mutations (results not shown).

The results for the reverse case, using the non-human-trained NN (15 parameters) are also very encouraging, with an overall success rate of 77 %, and a 52 % improvement over random (similar results were found for the two-parameter NN, see Table 2). Again, a more detailed look at the mutation-specific values show that NEMUs are less well predicted than expected. This is probably due to the same reasons exposed for the mouse-trained NN, that is: sample factors, a certain overtraining degree for the neutral mutations, and a slight bias originated by the selection procedure (see *Methods*).

We also explored whether other available methods could be utilised for the purpose of annotating animal model mutations. To this end we applied two well known prediction methods, SIFT [20] and PolyPhen [19] to the set of 105 mammalian DAMUs. We find that these two methods and our own method, give similar results: 73 %, 70 % and 72 % (75 %, if we consider only reliable predictions, as PolyPhen), respectively. Overall, these results confirm that human-based methods can be fruitfully applied to the annotation of animal model variability with reasonable success rates (as a reference of their expected behaviour, a comparison of the performance of the above methods when applied to score human mutations can be found in our previous work[25]).

In the present work we have focused in the study of DAMUs at the origin of Mendelian disorders. In the important case of polygenic diseases we expect disease-causing mutations to have smaller effects in the target protein, and therefore to be more difficult to identify. To assess to which extent our method may work scoring mutations associated to polygenic diseases, we tested our procedure in a set of over 4000 mutations in the *Escherichia coli* Lac repressor [56]. These mutations cover a broad range of functional effects, including very mild phenotypic effects. When applying our procedure to these data we find, as expected, a substantial decrease in the prediction performance: 63 % overall success rate, and 20 % improvement over random. Slightly better results are obtained utilising SIFT [20]: 68 % overall success rate, and 33 % improvement over random. However, it has to be noted that the Lac mutation dataset was utilised to parameterise SIFT [20], that is, the performance figures are an overestimation. Overall, both our results and their results indicate that prediction methods can be utilised to score mutations covering a broad range of phenotypic effects, including mild cases. However, the success rate will be clearly lower than that obtained for Mendelian DAMUs, and therefore care must be exercised. This obviously applies when transferring prediction tools to score mild mutations in other species.

CONCLUDING REMARKS

When considering protein sequence and structure, human and mouse are closely related organisms, and for this reason mutations leading to disease in one organism are likely to play the same role in the other. This suggests that human-based annotation tools could be utilised to fill the annotation gap in these and other species. Here we explore to which extent this can be done, using our neural network-based method [25]. We find that both human-trained NNs as well as mouse-trained NNs can be utilised to score mutations in the opposite species. In particular, we see that human-trained NNs can score with high confidence mouse DAMUs. However, a minimum sequence similarity among the sequences in both species (70 % - 80 %) guarantees that the prediction figures obtained in human will remain almost the same in mouse, or other species. Interestingly, the prediction process will work reasonably well even in the absence, from the training set, of human homologues of the mouse (or other species) proteins.

The fact that the smaller dataset –the mouse dataset- can be utilised to score a large number of human mutations with a good success rate, supports the strength of these results. Overall, our findings open the possibility to use mutation data collected from massive genomic experiments in animal models to predict the pathological profile of single point mutations in humans.

ACKNOWLEDGEMENTS

This work has been supported by the Spanish Ministry of Science and Technology (GEN2001-4758-C07-07, BIO2003-06848 and BIO2003-09327) and the Fundación Ramón Areces. The authors are grateful to Burkhardt Rost for lending us a copy of his program PHD, and Adrian Shepherd for lending us a copy of his FFNN neural networks package.

REFERENCES

1. Collins FS, Brooks LD, Chakravarti A. A DNA polymorphism discovery resource for research on human genetic variation. Genome Res 1998;8(12):1229-1231.
2. Chakravarti A. To a future of genetic medicine. Nature 2001;409(6822):822-823.
3. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, 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 JP, 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 JC, Mungall A, Plumb R, Ross M, Shownkeen R, Sims S, Waterston RH, Wilson RK, Hillier LW, McPherson JD, Marra MA, Mardis ER, Fulton LA, Chinwalla AT, Pepin KH, Gish WR, Chissoe SL, Wendl MC, Delehaunty KD, Miner TL, Delehaunty A, Kramer JB, Cook LL, Fulton RS, Johnson DL, Minx PJ, Clifton SW, Hawkins T, Branscomb E, Predki P, Richardson P, Wenning S, Slezak T, Doggett N, Cheng JF, Olsen A, Lucas S, Elkin C, Uberbacher E, Frazier M, Gibbs RA, Muzny DM, Scherer SE, Bouck JB, Sodergren EJ, Worley KC, Rives CM, Gorrell JH, Metzker ML, Naylor SL, Kucherlapati RS, Nelson DL, Weinstock GM, Sakaki Y, Fujiyama A, Hattori M, Yada T, Toyoda A, Itoh T, Kawagoe C, Watanabe H, Totoki Y, Taylor T, Weissenbach J, Heilig R, Saurin W, Artiguenave F, Brottier P, Bruls T, Pelletier E, Robert C, Wincker P, Smith DR, Doucette-Stamm L, Rubenfield M, Weinstock K, Lee HM, Dubois J, Rosenthal A, Platzer M, Nyakatura G, Taudien S, Rump A, Yang H, Yu J, Wang J, Huang G, Gu J, Hood L, Rowen L, Madan A, Qin S, Davis RW, Federspiel NA, Abola AP, Proctor MJ, Myers RM, Schmutz J, Dickson M, Grimwood J, Cox DR, Olson MV, Kaul R, Shimizu N, Kawasaki K, Minoshima S, Evans GA, Athanasiou M, Schultz R, Roe BA, Chen F, Pan H, Ramser J, Lehrach H, Reinhardt R, McCombie WR, de la Bastide M, Dedhia N, Blocker H, Hornischer K, Nordsiek G, Agarwala R, Aravind L, Bailey JA, Bateman A, Batzoglou S, Birney E, Bork P, Brown DG, Burge CB, Cerutti L, Chen HC, Church D, Clamp M, Copley RR, Doerks T, Eddy SR, Eichler EE, Furey TS, Galagan J, Gilbert JG, Harmon C, Hayashizaki Y, Haussler D, Hermjakob H, Hokamp K, Jang W, Johnson LS, Jones TA, Kasif S, Kaspryzk A, Kennedy S, Kent WJ, Kitts P, Koonin EV, Korf I, Kulp D, Lancet D, Lowe TM, McLysaght A, Mikkelsen T, Moran JV, Mulder N, Pollara VJ, Ponting CP, Schuler G, Schultz J, Slater G, Smit AF, Stupka E, Szustakowski J, Thierry-Mieg D, Thierry-Mieg J, Wagner L, Wallis J, Wheeler R, Williams A, Wolf YI, Wolfe KH, Yang SP, Yeh RF, Collins F, Guyer MS, Peterson J, Felsenfeld A, Wetterstrand KA, Patrinos A, Morgan MJ, Szustakowki J, de Jong P, Catanese JJ, Osoegawa K, Shizuya H, Choi S, Chen YJ. Initial sequencing and analysis of the human genome. Nature 2001;409(6822):860-921.

4. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, Gocayne JD, Amanatides P, Ballew RM, Huson DH, Wortman JR, Zhang Q, Kodira CD, Zheng XH, Chen L, Skupski M, Subramanian G, Thomas PD, Zhang J, Gabor Miklos GL, Nelson C, Broder S, Clark AG, Nadeau J, McKusick VA, Zinder N, Levine AJ, Roberts RJ, 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 AE, Gan W, Ge W, Gong F, Gu Z, Guan P, Heiman TJ, Higgins ME, Ji RR, Ke Z, Ketchum KA, Lai Z, Lei Y, Li Z, Li J, Liang Y, Lin X, Lu F, Merkulov GV, Milshina N, Moore HM, Naik AK, Narayan VA, Neelam B, Nusskern D, Rusch DB, Salzberg S, Shao W, Shue B, Sun J, Wang Z, Wang A, Wang X, Wang J, Wei M, Wides R, Xiao C, Yan C, Yao A, Ye J, Zhan M, Zhang W, Zhang H, Zhao Q, Zheng L, Zhong F, Zhong W, Zhu S, Zhao S, Gilbert D, Baumhueter S, Spier G, Carter C, Cravchik A, Woodage T, Ali F, An H, Awe A, Baldwin D, Baden H, Barnstead M, Barrow I, Beeson K, Busam D, Carver A, Center A, Cheng ML, Curry L, Danaher S, Davenport L, Desilets R, Dietz S, Dodson K, Doup L, Ferriera S, Garg N, Gluecksmann A, Hart B, Haynes J, Haynes C, Heiner C, Hladun S, Hostin D, Houck J, Howland T, Ibegwam C, Johnson J, Kalush F, Kline L, Koduru S, Love A, Mann F, May D, McCawley S, McIntosh T, McMullen I, Moy M, Moy L, Murphy B, Nelson K, Pfannkoch C, Pratts E, Puri V, Qureshi H, Reardon M, Rodriguez R, Rogers YH, Romblad D, Ruhfel B, Scott R, Sitter C, Smallwood M, Stewart E, Strong R, Suh E, Thomas R, Tint NN, Tse S, Vech C, Wang G, Wetter J, Williams S, Williams M, Windsor S, Winn-Deen E, Wolfe K, Zaveri J, Zaveri K, Abril JF, Guigo R, Campbell MJ, Sjolander KV, Karlak B, Kejariwal A, Mi H, Lazareva B, Hatton T, Narechania A, Diemer K, Muruganujan A, Guo N, Sato S, Bafna V, Istrail S, Lippert R, Schwartz R, Walenz B, Yooseph S, Allen D, Basu A, Baxendale J, Blick L, Caminha M, Carnes-Stine J, Caulk P, Chiang YH, Coyne M, Dahlke C, Mays A, Dombroski M, Donnelly M, Ely D, Esparham S, Fosler C, Gire H, Glanowski S, Glasser K, Glodek A, Gorokhov M, Graham K, Gropman B, Harris M, Heil J, Henderson S, Hoover J, Jennings D, Jordan C, Jordan J, Kasha J, Kagan L, Kraft C, Levitsky A, Lewis M, Liu X, Lopez J, Ma D, Majoros W, McDaniel J, Murphy S, Newman M, Nguyen T, Nguyen N, Nodell M, Pan S, Peck J, Peterson M, Rowe W, Sanders R, Scott J, Simpson M, Smith T, Sprague A, Stockwell T, Turner R, Venter E, Wang M, Wen M, Wu D, Wu M, Xia A, Zandieh A, Zhu X. The sequence of the human genome. Science 2001;291(5507):1304-1351.

5. Remm M, Metspalu A. High-density genotyping and linkage disequilibrium in the human genome using chromosome 22 as a model. Curr Opin Chem Biol 2002;6(1):24-30.

6. Sachidanandam R, Weissman D, Schmidt SC, Kakol JM, Stein LD, Marth G, Sherry S, Mullikin JC, Mortimore BJ, Willey DL, Hunt SE, Cole CG, Coggill

PC, Rice CM, Ning Z, Rogers J, Bentley DR, Kwok PY, Mardis ER, Yeh RT, Schultz B, Cook L, Davenport R, Dante M, Fulton L, Hillier L, Waterston RH, McPherson JD, Gilman B, Schaffner S, Van Etten WJ, Reich D, Higgins J, Daly MJ, Blumenstiel B, Baldwin J, Stange-Thomann N, Zody MC, Linton L, Lander ES, Altshuler D. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature 2001;409(6822):928-933.

7. Ferrer-Costa C, Orozco M, de la Cruz X. Characterization of Disease-associated Single Amino Acid Polymorphisms in Terms of Sequence and Structure Properties. J Mol Biol 2002;315(4):771-786.

8. Sunyaev S, Ramensky V, Koch I, Lathe W, 3rd, Kondrashov AS, Bork P. Prediction of deleterious human alleles. Hum Mol Genet 2001;10(6):591-597.

9. Rossant J, McKerlie C. Mouse-based phenogenomics for modelling human disease. Trends Mol Med 2001;7(11):502-507.

10. Erickson RP. Mouse models of human genetic disease: which mouse is more like a man? Bioessays 1996;18(12):993-998.

11. Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, An P, Antonarakis SE, Attwood J, Baertsch R, Bailey J, Barlow K, Beck S, Berry E, Birren B, Bloom T, Bork P, Botcherby M, Bray N, Brent MR, Brown DG, Brown SD, Bult C, Burton J, Butler J, Campbell RD, Carninci P, Cawley S, Chiaromonte F, Chinwalla AT, Church DM, Clamp M, Clee C, Collins FS, Cook LL, Copley RR, Coulson A, Couronne O, Cuff J, Curwen V, Cutts T, Daly M, David R, Davies J, Delehaunty KD, Deri J, Dermitzakis ET, Dewey C, Dickens NJ, Diekhans M, Dodge S, Dubchak I, Dunn DM, Eddy SR, Elnitski L, Emes RD, Eswara P, Eyras E, Felsenfeld A, Fewell GA, Flicek P, Foley K, Frankel WN, Fulton LA, Fulton RS, Furey TS, Gage D, Gibbs RA, Glusman G, Gnerre S, Goldman N, Goodstadt L, Grafham D, Graves TA, Green ED, Gregory S, Guigo R, Guyer M, Hardison RC, Haussler D, Hayashizaki Y, Hillier LW, Hinrichs A, Hlavina W, Holzer T, Hsu F, Hua A, Hubbard T, Hunt A, Jackson I, Jaffe DB, Johnson LS, Jones M, Jones TA, Joy A, Kamal M, Karlsson EK, Karolchik D, Kasprzyk A, Kawai J, Keibler E, Kells C, Kent WJ, Kirby A, Kolbe DL, Korf I, Kucherlapati RS, Kulbokas EJ, Kulp D, Landers T, Leger JP, Leonard S, Letunic I, Levine R, Li J, Li M, Lloyd C, Lucas S, Ma B, Maglott DR, Mardis ER, Matthews L, Mauceli E, Mayer JH, McCarthy M, McCombie WR, McLaren S, McLay K, McPherson JD, Meldrim J, Meredith B, Mesirov JP, Miller W, Miner TL, Mongin E, Montgomery KT, Morgan M, Mott R, Mullikin JC, Muzny DM, Nash WE, Nelson JO, Nhan MN, Nicol R, Ning Z, Nusbaum C, O'Connor MJ, Okazaki Y, Oliver K, Overton-Larty E, Pachter L, Parra G, Pepin KH, Peterson J, Pevzner P, Plumb R, Pohl CS, Poliakov A, Ponce TC, Ponting CP, Potter S, Quail M, Reymond A, Roe BA, Roskin KM, Rubin EM, Rust AG, Santos R, Sapojnikov V, Schultz B, Schultz J, Schwartz MS, Schwartz S, Scott C, Seaman S, Searle S, Sharpe T, Sheridan A, Shownkeen R, Sims S, Singer JB, Slater G, Smit A, Smith DR, Spencer B, Stabenau A, Stange-Thomann N, Sugnet C, Suyama M, Tesler G, Thompson J, Torrents D, Trevaskis E, Tromp J, Ucla C, Ureta-Vidal A, Vinson JP, Von Niederhausern AC, Wade CM, Wall M, Weber RJ, Weiss RB, Wendl MC, West AP, Wetterstrand K, Wheeler R, Whelan S, Wierzbowski J, Willey D, Williams S, Wilson RK, Winter E, Worley KC, Wyman D, Yang S, Yang SP, Zdobnov EM, Zody MC, Lander ES. Initial

sequencing and comparative analysis of the mouse genome. Nature 2002;420(6915):520-562.

12. Russell RB, Barton GJ. Structural features can be unconserved in proteins with similar folds. An analysis of side-chain to side-chain contacts secondary structure and accessibility. J Mol Biol 1994;244(3):332-350.

13. Hrabe de Angelis MH, Flaswinkel H, Fuchs H, Rathkolb B, Soewarto D, Marschall S, Heffner S, Pargent W, Wuensch K, Jung M, Reis A, Richter T, Alessandrini F, Jakob T, Fuchs E, Kolb H, Kremmer E, Schaeble K, Rollinski B, Roscher A, Peters C, Meitinger T, Strom T, Steckler T, Holsboer F, Klopstock T, Gekeler F, Schindewolf C, Jung T, Avraham K, Behrendt H, Ring J, Zimmer A, Schughart K, Pfeffer K, Wolf E, Balling R. Genome-wide, large-scale production of mutant mice by ENU mutagenesis. Nat Genet 2000;25(4):444-447.

14. Nolan PM, Peters J, Strivens M, Rogers D, Hagan J, Spurr N, Gray IC, Vizor L, Brooker D, Whitehill E, Washbourne R, Hough T, Greenaway S, Hewitt M, Liu X, McCormack S, Pickford K, Selley R, Wells C, Tymowska-Lalanne Z, Roby P, Glenister P, Thornton C, Thaung C, Stevenson JA, Arkell R, Mburu P, Hardisty R, Kiernan A, Erven A, Steel KP, Voegeling S, Guenet JL, Nickols C, Sadri R, Nasse M, Isaacs A, Davies K, Browne M, Fisher EM, Martin J, Rastan S, Brown SD, Hunter J. A systematic, genome-wide, phenotype-driven mutagenesis programme for gene function studies in the mouse. Nat Genet 2000;25(4):440-443.

15. Gibbs RA, Weinstock GM, Metzker ML, Muzny DM, Sodergren EJ, Scherer S, Scott G, Steffen D, Worley KC, Burch PE, Okwuonu G, Hines S, Lewis L, DeRamo C, Delgado O, Dugan-Rocha S, Miner G, Morgan M, Hawes A, Gill R, Celera, Holt RA, Adams MD, Amanatides PG, Baden-Tillson H, Barnstead M, Chin S, Evans CA, Ferriera S, Fosler C, Glodek A, Gu Z, Jennings D, Kraft CL, Nguyen T, Pfannkoch CM, Sitter C, Sutton GG, Venter JC, Woodage T, Smith D, Lee HM, Gustafson E, Cahill P, Kana A, Doucette-Stamm L, Weinstock K, Fechtel K, Weiss RB, Dunn DM, Green ED, Blakesley RW, Bouffard GG, De Jong PJ, Osoegawa K, Zhu B, Marra M, Schein J, Bosdet I, Fjell C, Jones S, Krzywinski M, Mathewson C, Siddiqui A, Wye N, McPherson J, Zhao S, Fraser CM, Shetty J, Shatsman S, Geer K, Chen Y, Abramzon S, Nierman WC, Havlak PH, Chen R, Durbin KJ, Egan A, Ren Y, Song XZ, Li B, Liu Y, Qin X, Cawley S, Cooney AJ, D'Souza LM, Martin K, Wu JQ, Gonzalez-Garay ML, Jackson AR, Kalafus KJ, McLeod MP, Milosavljevic A, Virk D, Volkov A, Wheeler DA, Zhang Z, Bailey JA, Eichler EE, Tuzun E, Birney E, Mongin E, Ureta-Vidal A, Woodwark C, Zdobnov E, Bork P, Suyama M, Torrents D, Alexandersson M, Trask BJ, Young JM, Huang H, Wang H, Xing H, Daniels S, Gietzen D, Schmidt J, Stevens K, Vitt U, Wingrove J, Camara F, Mar Alba M, Abril JF, Guigo R, Smit A, Dubchak I, Rubin EM, Couronne O, Poliakov A, Hubner N, Ganten D, Goesele C, Hummel O, Kreitler T, Lee YA, Monti J, Schulz H, Zimdahl H, Himmelbauer H, Lehrach H, Jacob HJ, Bromberg S, Gullings-Handley J, Jensen-Seaman MI, Kwitek AE, Lazar J, Pasko D, Tonellato PJ, Twigger S, Ponting CP, Duarte JM, Rice S, Goodstadt L, Beatson

SA, Emes RD, Winter EE, Webber C, Brandt P, Nyakatura G, Adetobi M, Chiaromonte F, Elnitski L, Eswara P, Hardison RC, Hou M, Kolbe D, Makova K, Miller W, Nekrutenko A, Riemer C, Schwartz S, Taylor J, Yang S, Zhang Y, Lindpaintner K, Andrews TD, Caccamo M, Clamp M, Clarke L, Curwen V, Durbin R, Eyras E, Searle SM, Cooper GM, Batzoglou S, Brudno M, Sidow A, Stone EA, Payseur BA, Bourque G, Lopez-Otin C, Puente XS, Chakrabarti K, Chatterji S, Dewey C, Pachter L, Bray N, Yap VB, Caspi A, Tesler G, Pevzner PA, Haussler D, Roskin KM, Baertsch R, Clawson H, Furey TS, Hinrichs AS, Karolchik D, Kent WJ, Rosenbloom KR, Trumbower H, Weirauch M, Cooper DN, Stenson PD, Ma B, Brent M, Arumugam M, Shteynberg D, Copley RR, Taylor MS, Riethman H, Mudunuri U, Peterson J, Guyer M, Felsenfeld A, Old S, Mockrin S, Collins F. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. Nature 2004;428(6982):493-521.

16. Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG, Scherer SE, Li PW, Hoskins RA, Galle RF, George RA, Lewis SE, Richards S, Ashburner M, Henderson SN, Sutton GG, Wortman JR, Yandell MD, Zhang Q, Chen LX, Brandon RC, Rogers YH, Blazej RG, Champe M, Pfeiffer BD, Wan KH, Doyle C, Baxter EG, Helt G, Nelson CR, Gabor GL, Abril JF, Agbayani A, An HJ, Andrews-Pfannkoch C, Baldwin D, Ballew RM, Basu A, Baxendale J, Bayraktaroglu L, Beasley EM, Beeson KY, Benos PV, Berman BP, Bhandari D, Bolshakov S, Borkova D, Botchan MR, Bouck J, Brokstein P, Brottier P, Burtis KC, Busam DA, Butler H, Cadieu E, Center A, Chandra I, Cherry JM, Cawley S, Dahlke C, Davenport LB, Davies P, de Pablos B, Delcher A, Deng Z, Mays AD, Dew I, Dietz SM, Dodson K, Doup LE, Downes M, Dugan-Rocha S, Dunkov BC, Dunn P, Durbin KJ, Evangelista CC, Ferraz C, Ferriera S, Fleischmann W, Fosler C, Gabrielian AE, Garg NS, Gelbart WM, Glasser K, Glodek A, Gong F, Gorrell JH, Gu Z, Guan P, Harris M, Harris NL, Harvey D, Heiman TJ, Hernandez JR, Houck J, Hostin D, Houston KA, Howland TJ, Wei MH, Ibegwam C, Jalali M, Kalush F, Karpen GH, Ke Z, Kennison JA, Ketchum KA, Kimmel BE, Kodira CD, Kraft C, Kravitz S, Kulp D, Lai Z, Lasko P, Lei Y, Levitsky AA, Li J, Li Z, Liang Y, Lin X, Liu X, Mattei B, McIntosh TC, McLeod MP, McPherson D, Merkulov G, Milshina NV, Mobarry C, Morris J, Moshrefi A, Mount SM, Moy M, Murphy B, Murphy L, Muzny DM, Nelson DL, Nelson DR, Nelson KA, Nixon K, Nusskern DR, Pacleb JM, Palazzolo M, Pittman GS, Pan S, Pollard J, Puri V, Reese MG, Reinert K, Remington K, Saunders RD, Scheeler F, Shen H, Shue BC, Siden-Kiamos I, Simpson M, Skupski MP, Smith T, Spier E, Spradling AC, Stapleton M, Strong R, Sun E, Svirskas R, Tector C, Turner R, Venter E, Wang AH, Wang X, Wang ZY, Wassarman DA, Weinstock GM, Weissenbach J, Williams SM, WoodageT, Worley KC, Wu D, Yang S, Yao QA, Ye J, Yeh RF, Zaveri JS, Zhan M, Zhang G, Zhao Q, Zheng L, Zheng XH, Zhong FN, Zhong W, Zhou X, Zhu S, Zhu X, Smith HO, Gibbs RA, Myers EW, Rubin GM, Venter JC. The genome sequence of Drosophila melanogaster. Science 2000;287(5461):2185-2195.

17. Celniker SE, Wheeler DA, Kronmiller B, Carlson JW, Halpern A, Patel S, Adams M, Champe M, Dugan SP, Frise E, Hodgson A, George RA, Hoskins RA, Laverty T, Muzny DM, Nelson CR, Pacleb JM, Park S, Pfeiffer BD, Richards S, Sodergren EJ, Svirskas R, Tabor PE, Wan K, Stapleton M, Sutton GG, Venter C, Weinstock G, Scherer SE, Myers EW, Gibbs RA, Rubin GM.

Finishing a whole-genome shotgun: release 3 of the Drosophila melanogaster euchromatic genome sequence. Genome Biol 2002;3(12):RESEARCH0079.

18. Guryev V, Berezikov E, Malik R, Plasterk RH, Cuppen E. Single nucleotide polymorphisms associated with rat expressed sequences. Genome Res 2004;14(7):1438-1443.
19. Ramensky V, Bork P, Sunyaev S. Human non-synonymous SNPs: server and survey. Nucleic Acids Res 2002;30(17):3894-3900.
20. Ng PC, Henikoff S. Predicting deleterious amino acid substitutions. Genome Res 2001;11(5):863-874.
21. Chasman D, Adams RM. Predicting the functional consequences of non-synonymous single nucleotide polymorphisms: structure-based assessment of amino acid variation. J Mol Biol 2001;307(2):683-706.
22. Conde L, Vaquerizas JM, Santoyo J, Al-Shahrour F, Ruiz-Llorente S, Robledo M, Dopazo J. PupaSNP Finder: a web tool for finding SNPs with putative effect at transcriptional level. Nucleic Acids Res 2004;32(Web Server issue):W242-248.
23. Santibanez-Koref MF, Gangeswaran R, Santibanez-Koref IP, Shanahan N, Hancock JM. A phylogenetic approach to assessing the significance of missense mutations in disease genes. Hum Mutat 2003;22(1):51-58.
24. Wang Z, Moult J. SNPs, protein structure, and disease. Hum Mutat 2001;17(4):263-270.
25. Ferrer-Costa C, Orozco M, De La Cruz X. Sequence-based prediction of pathological mutations. Proteins 2004;57:811-819.
26. Livingston RJ, von Niederhausern A, Jegga AG, Crawford DC, Carlson CS, Rieder MJ, Gowrisankar S, Aronow BJ, Weiss RB, Nickerson DA. Pattern of sequence variation across 213 environmental response genes. Genome Res 2004;14(10):1821-1831.
27. Xi T, Jones IM, Mohrenweiser HW. Many amino acid substitution variants identified in DNA repair genes during human population screenings are predicted to impact protein function. Genomics 2004;83(6):970-979.
28. Krishnan VG, Westhead DR. A comparative study of machine-learning methods to predict the effects of single nucleotide polymorphisms on protein function. Bioinformatics 2003;19(17):2199-2209.
29. de La Cruz X, Calvo M. Use of surface area computations to describe atom-atom interactions. J Comput Aided Mol Des 2001;15(6):521-532.
30. Martin AC, Facchiano AM, Cuff AL, Hernandez-Boussard T, Olivier M, Hainaut P, Thornton JM. Integrating mutation data and structural analysis of the TP53 tumor-suppressor protein. Hum Mutat 2002;19(2):149-164.
31. Steward RE, MacArthur MW, Laskowski RA, Thornton JM. Molecular basis of inherited diseases: a structural perspective. Trends Genet 2003;19(9):505-513.
32. Zuckerkandl E, Pauling L. Molecular disease, evolution and genetic heterogeneity. In: B. M, Pullman B, editors. Horizons in Biochemisty. London: Academic Press; 1962. p 189-225.

33. Kondrashov AS, Sunyaev S, Kondrashov FA. Dobzhansky-Muller incompatibilities in protein evolution. Proc Natl Acad Sci U S A 2002;99(23):14878-14883.
34. Huang H, Winter EE, Wang H, Weinstock KG, Xing H, Goodstadt L, Stenson PD, Cooper DN, Smith D, Alba MM, Ponting CP, Fechtel K. Evolutionary conservation and selection of human disease gene orthologs in the rat and mouse genomes. Genome Biol 2004;5(7):R47.
35. Gao L, Zhang J. Why are some human disease-associated mutations fixed in mice? Trends Genet 2003;19(12):678-681.
36. Rost B, Sander C. Prediction of protein secondary structure at better than 70% accuracy. J Mol Biol 1993;232(2):584-599.
37. Fauchere JL, Pliska V. Hydrophobic parameters of amino acid side chains from the partitioning of N-acetyl-amino acid amides. European Journal of Medicinal Chemistry 1983;18(4):369-375.
38. Miller S, Lesk AM, Janin J, Chothia C. The accessible surface area and stability of oligomeric proteins. Nature 1987;328(6133):834-836.
39. Chou PY, Fasman GD. Conformational parameters for amino acids in helical, beta-sheet, and random coil regions calculated from proteins. Biochemistry 1974;13(2):211-222.
40. Swindells MB, MacArthur MW, Thornton JM. Intrinsic phi, psi propensities of amino acids, derived from the coil regions of known structures. Nat Struct Biol 1995;2(7):596-603.
41. Bondi A. Van der Waals volumes and radii. J Phys Chem 1964;68(3):441-451.
42. Chothia C. Structural invariants in protein folding. Nature 1975;254(5498):304-308.
43. Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci U S A 1992;89(22):10915-10919.
44. Dayhoff MO, Schwartz, R. M. & Orcutt, B. C. A model of evolutionary change in proteins. In: Dayhoff MO, editor. Atlas of protein sequence and structure. Washington D. C.: National Biomedical Research Foundation; 1978. p 345-352.
45. Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, Martin MJ, Natale DA, O'Donovan C, Redaschi N, Yeh LS. UniProt: the Universal Protein knowledgebase. Nucleic Acids Res 2004;32 Database issue:D115-119.
46. Bateman A, Birney E, Cerruti L, Durbin R, Etwiller L, Eddy SR, Griffiths-Jones S, Howe KL, Marshall M, Sonnhammer EL. The Pfam protein families database. Nucleic Acids Res 2002;30(1):276-280.
47. Shannon CE. A mathematical theory of communication. Bell System Tech J 1948;27:379-423.
48. Sippl MJ. Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins. J Mol Biol 1990;213(4):859-883.
49. Goodstadt L, Ponting CP. Sequence variation and disease in the wake of the draft human genome. Hum Mol Genet 2001;10(20):2209-2214.
50. Harris MA, Clark J, Ireland A, Lomax J, Ashburner M, Foulger R, Eilbeck K, Lewis S, Marshall B, Mungall C, Richter J, Rubin GM, Blake JA, Bult C, Dolan M, Drabkin H, Eppig JT, Hill DP, Ni L, Ringwald M, Balakrishnan R, Cherry JM, Christie KR, Costanzo MC, Dwight SS, Engel S, Fisk DG, Hirschman JE,

Hong EL, Nash RS, Sethuraman A, Theesfeld CL, Botstein D, Dolinski K, Feierbach B, Berardini T, Mundodi S, Rhee SY, Apweiler R, Barrell D, Camon E, Dimmer E, Lee V, Chisholm R, Gaudet P, Kibbe W, Kishore R, Schwarz EM, Sternberg P, Gwinn M, Hannick L, Wortman J, Berriman M, Wood V, de la Cruz N, Tonellato P, Jaiswal P, Seigfried T, White R. The Gene Ontology (GO) database and informatics resource. Nucleic Acids Res 2004;32 Database issue:D258-261.

51. Al-Shahrour F, Diaz-Uriarte R, Dopazo J. FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes. Bioinformatics 2004;20(4):578-580.
52. Rumelhart DE, McClelland JL, University of California San Diego. PDP Research Group. Parallel distributed processing : explorations in the microstructure of cognition. Cambridge, Mass.: MIT Press; 1986. 2 v. p.
53. Mehrotra K, Mohan CK, Ranka S, NetLibrary Inc. Elements of artificial neural networks. Cambridge, Mass.: MIT Press; 1997. xiv, 344 p.
54. Shepherd AJ, Gorse D, Thornton JM. Prediction of the location and type of beta-turns in proteins using neural networks. Protein Sci 1999;8(5):1045-1055.
55. Lopez-Bigas N, Ouzounis CA. Genome-wide identification of genes likely to be involved in human genetic disease. Nucleic Acids Res 2004;32(10):3108-3114.
56. Suckow J, Markiewicz P, Kleina LG, Miller J, Kisters-Woike B, Muller-Hill B. Genetic studies of the Lac repressor. XV: 4000 single amino acid substitutions and analysis of the resulting phenotypes on the basis of the protein structure. J Mol Biol 1996;261(4):509-523.
57. Press WH, Flannery BP, Teukolsky SA, Vetterling WT. Numerical recipes in C : the art of scientific computing. Cambridge ; New York: Cambridge University Press; 1992. xxvi, 994 p.

LEGENDS TO FIGURES

Figure 1. Distribution of sequence identities between: (i) human proteins and mouse SwissProt [45] homologues (BLACK); (ii) human protein and mouse proteins, in which both are associated to disease, according to SwissProt annotations (HASHED); and (iii) human proteins and non-human homologues, where the latter carry in, at least one position, one amino acid that would be pathological in human (GREY).

Figure 2. Distribution of the level-2 gene ontology terms for the set of mouse proteins carrying at least one DAMU, and for the whole set of mouse sequences in the SwissProt database [45]. As mentioned in the text, no significant differences were found among both datasets.

Figure 3. Comparison between mouse (RED) and human (YELLOW) distributions for selected descriptors for pathological mutations. The chi-square tests done to compare them indicate that there is no significant difference among both species for these properties: accessibility, p-value=0.608; hydrophobicity, p-value=1.000; Blosum62, p-value=0.998; PSSM, p-value=1.000; DPSSM, p-value=1.000; Shannon's Entropy, p-value=1.000.

Table 1. Performance measures (see *Methods*) of the NN trained to differentiate between human and non-human DAMUs (average plus standard deviations are shown). Values were obtained using 15 descriptors (see *Methods*). All refers to the non-human dataset in which other animal mutations were added to the mouse dataset (see *Methods*).

Descriptor	mouse	all
Q_{tot}	90.7 ± 1.7	92.21±0.20
S_{tot}	0.5 ± 0.7	1.1±2.7
S^{mouse}	0.4 ± 0.6	0.7±1.8
S^{human}	0.6 ± 0.9	2.1±6.4

Table 2. Parameters describing the prediction performance in mouse mutations of human-trained NNs (column 2), and *vice versa* (column 3). The NNs were trained with our human mutations set (see Method. Values in BOLD correspond to results obtained using 15 descriptors, values in ITALICS to those obtained using two 2 descriptors (PSSM and residue hydrophobicities). Also shown the performances of the human-trained NN when applied to the all species dataset (column 4), and those of the reverse problem (column 5). In columns 4 and 5, ALL refers to the non-human mutation dataset in which mouse and other animal mutations were put together (see *Methods*).

	HUMAN TO MOUSE	MOUSE TO HUMAN	HUMAN TO ALL	ALL TO HUMAN
Qtot	**86** 86	**78** 77	**85** 89	**77** 76
S_{tot}	**53** 51	**55** 51	**42** 51	**52** 49
S^{NEMU}	**49** 49	**69** 70	**31** 43	**69** 70
S^{DAMU}	**60** 53	**46** 41	**66** 63	**41** 38

FIGURE 1

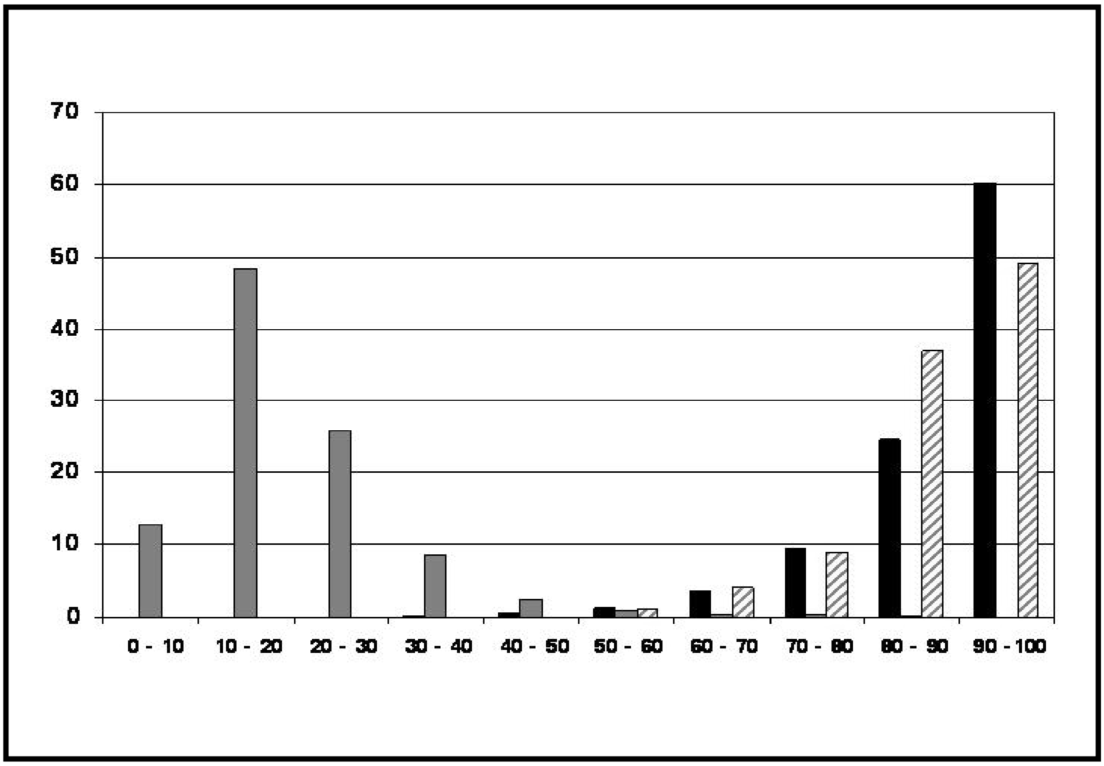

70
60
50
40
30
20
10
0
0 - 10
10 - 20
20 - 30
30 - 40
40 - 50
50 - 60
60 - 70
70 - 80
80 - 90
90 - 100

FIGURE 2

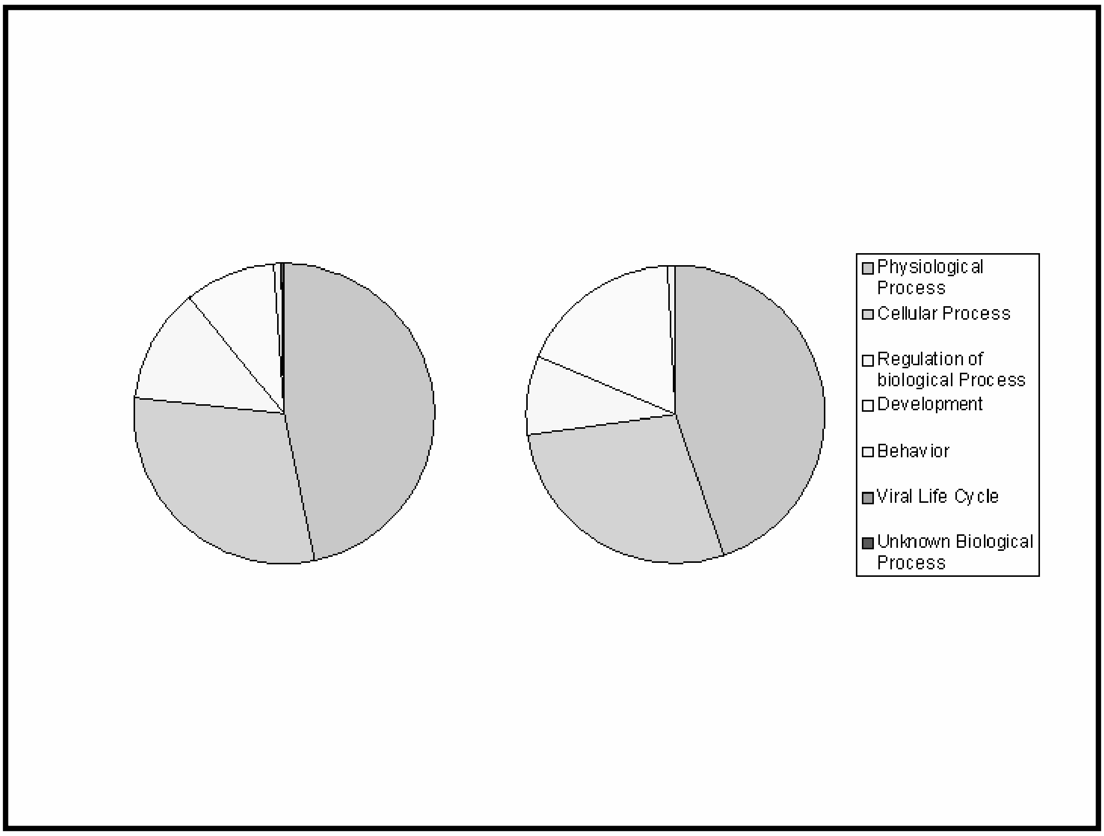
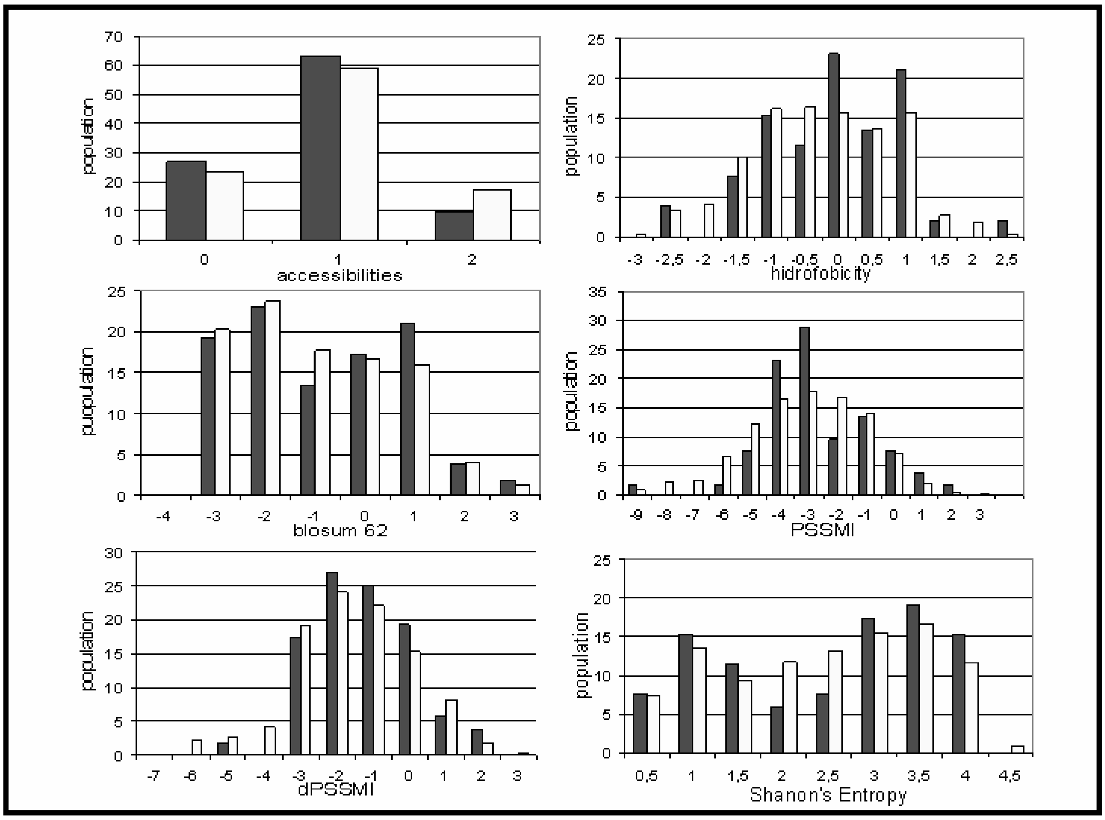

Physiological Process
Cellular Process
Regulation of biological Process
Development
Behavior
Viral Life Cycle
Unknown Biological Process

FIGURE 3


population
0
10
20
30
40
50
60
70
0
1
2
accessibilities
population
0
5
10
15
20
25
-3 -2,5 -2 -1,5 -1 -0,5 0 0,5 1 1,5 2 2,5
hidrofobicity
pupulation
0
5
10
15
20
25
-4 -3 -2 -1 0 1 2 3
blosum 62
population
0
5
10
15
20
25
30
35
-9 -8 -7 -6 -5 -4 -3 -2 -1 0 1 2 3
PSSMI
population
0
5
10
15
20
25
30
-7 -6 -5 -4 -3 -2 -1 0 1 2 3
dPSSMI
population
0
5
10
15
20
25
0,5 1 1,5 2 2,5 3 3,5 4 4,5
Shanon's Entropy

APPENDIX

In the present work we explore to which extent tools developed to predict DAMUs in one species, e.g. human, can be transferred to another species, e.g. mouse. To this end a NN is trained with a set of mutations from one of the species and tested in set of mutations from the other species, analysing the performance of the trained NN using standard measures like Q_{tot} or S_{tot}. The use of these performance measures must be done with caution, since for finite training/test sets values can oscillate due to factors like sample size and proportion of neutral/pathological mutations in the training and test sets. Thus, the significance of performance measures must be always analysed considering the characteristics of the test sample. To this end we have developed a simple probabilistic background model. Deviations from this model might signal the presence of some biasing factors, like actual differences among species, or technical effects of the NN training procedure, which can compromise the validity of the performance measures.

For each mutation type, the NN scoring process of the corresponding mutation set can be described by a binomial distribution with parameters N and q, where N is the total number of mutations, and q is the probability of a correct prediction. For example, N can be the number of mouse DAMUs, and q the probability of correctly predicting them. As an estimation of the later we will use the value obtained for the training set. For example, when analysing the results of the prediction of mouse DAMUs with the human-trained NN we took: N=75 DAMUs (the total number of mouse DAMUs) and q=0.73 (the prediction success for human DAMUs). For the mouse NEMUs, we took N=373 (the total number of mouse NEMUs) and q= 0.83 (the prediction success for human NEMUs).

We applied this simple probabilistic model to compute the probability density of S^{DAMU} and S^{NEMU} (two magnitudes that provide a fair view of the performance of the NN). Note that according to eqs. 8a-8b (see methods) S^{DAMU} and S^{NEMU} are functions of four parameters: p, n, o and u. p and u are the number of correctly and incorrectly predicted DAMUs, respectively; n and o are the number of correctly and incorrectly predicted NEMUs, respectively. Note that $u+p=N_d$ and $N_u=n+o$, where N_d and N_n are the total number of DAMUs and NEMUs in the test set. According to our background probabilistic model, p and n (in the test set) should follow a binomial distribution, with parameters (N_d, q_d) and (N_n, q_n), respectively, where q_d and q_n are the probabilities that the NN successfully predicts DAMUs and NEMUs (estimated from the training set).

Once a binomial distribution for p and n are generated -using the corresponding routine in "Numerical Recipes in C" [57] to generate a set of 10^6 values of p and n, and the associated o and u- we computed the background distribution of S^{DAMU} and S^{NEMU} (see Figure 1 in Appendix) that should be expected in a test set for a given NN performance (determined by q_d and q_n), composition and size (determined by N_d and N_n).

Values of the real S^{DAMU} and S^{NEMU} found when the test set was analysed using the trained NN should be then compared with those computed by the theoretical binomial distribution for the same test sample. Large deviations from the average values point to training problems arising from fundamental differences between the samples, from overtraining, or from large imbalances between DAMUs and NEMUs in the training set.

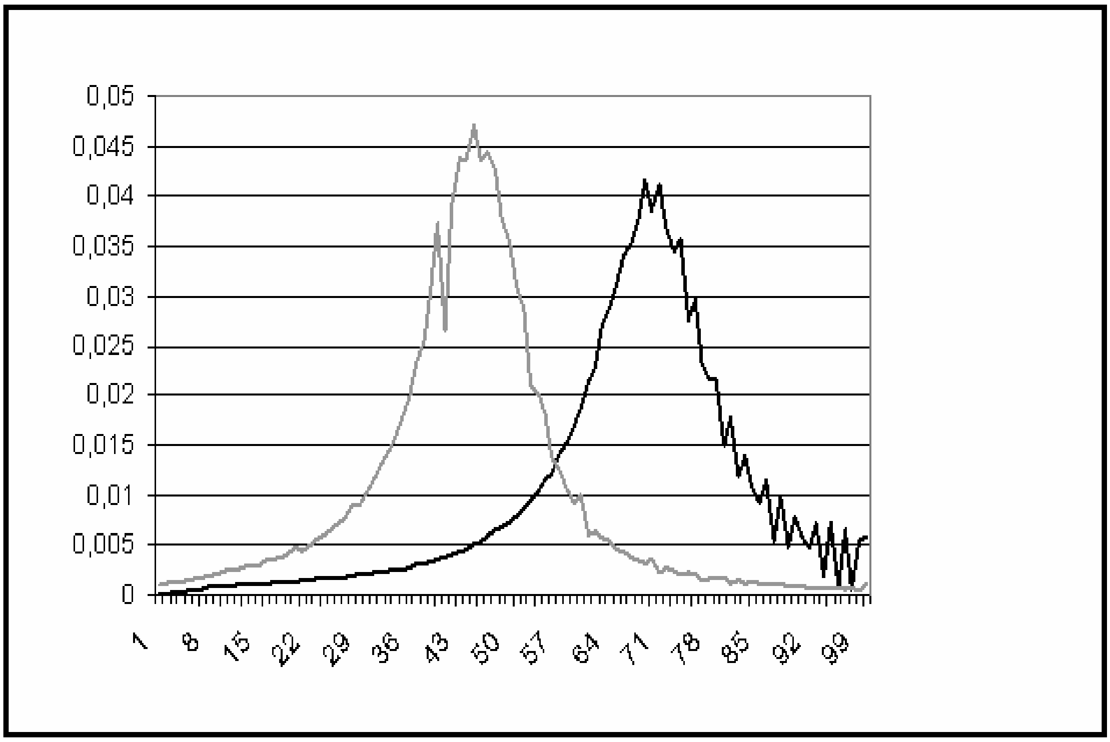
REFERENCES

1. Press WH, Flannery BP, Teukolsky SA, Vetterling WT. Numerical recipes in C : the art of scientific computing. Cambridge ; New York: Cambridge University Press; 1992. xxvi, 994 p

FIGURE CAPTIONS APPENDIX

FIGURE 1. Simulated probability density for S^{DAMU} (BLACK) and S^{NEMU} (DARK GREY) when utilising the human-trained NN to predict the mouse set (see Appendix text).

FIGURE 1. APPENDIX.



IV. BIBLIOGRAFIA DEL CAPÍTOL

Adams, M. D., Celniker, S. E., Holt, R. A., Evans, C. A., Gocayne, J. D., Amanatides, P. G., Scherer, S. E., Li, P. W., Hoskins, R. A., Galle, R. F., George, R. A., Lewis, S. E., Richards, S., Ashburner, M., Henderson, S. N., Sutton, G. G., Wortman, J. R., Yandell, M. D., Zhang, Q., Chen, L. X., Brandon, R. C., Rogers, Y. H., Blazej, R. G., Champe, M., Pfeiffer, B. D., Wan, K. H., Doyle, C., Baxter, E. G., Helt, G., Nelson, C. R., Gabor, G. L., Abril, J. F., Agbayani, A., An, H. J., Andrews-Pfannkoch, C., Baldwin, D., Ballew, R. M., Basu, A., Baxendale, J., Bayraktaroglu, L., Beasley, E. M., Beeson, K. Y., Benos, P. V., Berman, B. P., Bhandari, D., Bolshakov, S., Borkova, D., Botchan, M. R., Bouck, J., Brokstein, P., Brottier, P., Burtis, K. C., Busam, D. A., Butler, H., Cadieu, E., Center, A., Chandra, I., Cherry, J. M., Cawley, S., Dahlke, C., Davenport, L. B., Davies, P., de Pablos, B., Delcher, A., Deng, Z., Mays, A. D., Dew, I., Dietz, S. M., Dodson, K., Doup, L. E., Downes, M., Dugan-Rocha, S., Dunkov, B. C., Dunn, P., Durbin, K. J., Evangelista, C. C., Ferraz, C., Ferriera, S., Fleischmann, W., Fosler, C., Gabrielian, A. E., Garg, N. S., Gelbart, W. M., Glasser, K., Glodek, A., Gong, F., Gorrell, J. H., Gu, Z., Guan, P., Harris, M., Harris, N. L., Harvey, D., Heiman, T. J., Hernandez, J. R., Houck, J., Hostin, D., Houston, K. A., Howland, T. J., Wei, M. H., Ibegwam, C., et al. (2000). The genome sequence of Drosophila melanogaster. *Science* 287, 2185-95.

Al-Shahrour, F., Diaz-Uriarte, R. & Dopazo, J. (2004). FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes. *Bioinformatics* 20, 578-80.

Apweiler, R., Bairoch, A., Wu, C. H., Barker, W. C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., Martin, M. J., Natale, D. A., O'Donovan, C., Redaschi, N. & Yeh, L. S. (2004). UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res* 32 Database issue, D115-9.

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.

Celniker, S. E., Wheeler, D. A., Kronmiller, B., Carlson, J. W., Halpern, A., Patel, S., Adams, M., Champe, M., Dugan, S. P., Frise, E., Hodgson, A., George, R. A., Hoskins, R. A., Laverty, T., Muzny, D. M., Nelson, C. R., Pacleb, J. M., Park, S., Pfeiffer, B. D., Richards, S., Sodergren, E. J., Svirskas, R., Tabor, P. E., Wan, K., Stapleton, M., Sutton, G. G., Venter, C., Weinstock, G., Scherer, S. E., Myers, E. W., Gibbs, R. A. &

Rubin, G. M. (2002). Finishing a whole-genome shotgun: release 3 of the Drosophila melanogaster euchromatic genome sequence. *Genome Biol* 3, RESEARCH0079.

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.

Conde, L., Vaquerizas, J. M., Santoyo, J., Al-Shahrour, F., Ruiz-Llorente, S., Robledo, M. & Dopazo, J. (2004). PupaSNP Finder: a web tool for finding SNPs with putative effect at transcriptional level. *Nucleic Acids Res* 32, W242-8.

Fauchere, J. L. & Pliska, V. (1983). Hydrophobic parameters of amino acid side chains from the partitioning of N-acetyl-amino acid amides. *European Journal of Medicinal Chemistry* 18, 369-75.

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.

Gibbs, R. A., Weinstock, G. M., Metzker, M. L., Muzny, D. M., Sodergren, E. J., Scherer, S., Scott, G., Steffen, D., Worley, K. C., Burch, P. E., Okwuonu, G., Hines, S., Lewis, L., DeRamo, C., Delgado, O., Dugan-Rocha, S., Miner, G., Morgan, M., Hawes, A., Gill, R., Celera, Holt, R. A., Adams, M. D., Amanatides, P. G., Baden-Tillson, H., Barnstead, M., Chin, S., Evans, C. A., Ferriera, S., Fosler, C., Glodek, A., Gu, Z., Jennings, D., Kraft, C. L., Nguyen, T., Pfannkoch, C. M., Sitter, C., Sutton, G. G., Venter, J. C., Woodage, T., Smith, D., Lee, H. M., Gustafson, E., Cahill, P., Kana, A., Doucette-Stamm, L., Weinstock, K., Fechtel, K., Weiss, R. B., Dunn, D. M., Green, E. D., Blakesley, R. W., Bouffard, G. G., De Jong, P. J., Osoegawa, K., Zhu, B., Marra, M., Schein, J., Bosdet, I., Fjell, C., Jones, S., Krzywinski, M., Mathewson, C., Siddiqui, A., Wye, N., McPherson, J., Zhao, S., Fraser, C. M., Shetty, J., Shatsman, S., Geer, K., Chen, Y., Abramzon, S., Nierman, W. C., Havlak, P. H., Chen, R., Durbin, K. J., Egan, A., Ren, Y., Song, X. Z., Li, B., Liu, Y., Qin, X., Cawley, S., Cooney, A. J., D'Souza, L. M., Martin, K., Wu, J. Q., Gonzalez-Garay, M. L., Jackson, A. R., Kalafus, K. J., McLeod, M. P., Milosavljevic, A., Virk, D., Volkov, A., Wheeler, D. A., Zhang, Z., Bailey, J. A., Eichler, E. E., Tuzun, E., et al. (2004). Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* 428, 493-521.

Goodstadt, L. & Ponting, C. P. (2001). Sequence variation and disease in the wake of the draft human genome. *Hum Mol Genet* 10, 2209-14.

Harris, M. A., Clark, J., Ireland, A., Lomax, J., Ashburner, M., Foulger, R., Eilbeck, K., Lewis, S., Marshall, B., Mungall, C., Richter, J., Rubin, G. M., Blake, J. A., Bult, C., Dolan, M., Drabkin, H., Eppig, J. T., Hill, D. P., Ni, L., Ringwald, M., Balakrishnan, R., Cherry, J. M., Christie, K. R., Costanzo, M. C., Dwight, S. S., Engel, S., Fisk, D. G., Hirschman, J. E., Hong, E. L., Nash, R. S., Sethuraman, A., Theesfeld, C. L., Botstein, D., Dolinski, K., Feierbach, B., Berardini, T., Mundodi, S., Rhee, S. Y., Apweiler, R., Barrell, D., Camon, E., Dimmer, E., Lee, V., Chisholm, R., Gaudet, P., Kibbe, W., Kishore, R., Schwarz, E. M., Sternberg, P., Gwinn, M., Hannick, L., Wortman, J., Berriman, M., Wood, V., de la Cruz, N., Tonellato, P., Jaiswal, P., Seigfried, T. & White, R. (2004). The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res* 32 Database issue, D258-61.

Hrabe de Angelis, M. H., Flaswinkel, H., Fuchs, H., Rathkolb, B., Soewarto, D., Marschall, S., Heffner, S., Pargent, W., Wuensch, K., Jung, M., Reis, A., Richter, T., Alessandrini, F., Jakob, T., Fuchs, E., Kolb, H., Kremmer, E., Schaeble, K., Rollinski, B., Roscher, A., Peters, C., Meitinger, T., Strom, T., Steckler, T., Holsboer, F., Klopstock, T., Gekeler, F., Schindewolf, C., Jung, T., Avraham, K., Behrendt, H., Ring, J., Zimmer, A., Schughart, K., Pfeffer, K., Wolf, E. & Balling, R. (2000). Genome-wide, large-scale production of mutant mice by ENU mutagenesis. *Nat Genet* 25, 444-7.

Huang, H., Winter, E. E., Wang, H., Weinstock, K. G., Xing, H., Goodstadt, L., Stenson, P. D., Cooper, D. N., Smith, D., Alba, M. M., Ponting, C. P. & Fechtel, K. (2004). Evolutionary conservation and selection of human disease gene orthologs in the rat and mouse genomes. *Genome Biol* 5, R47.

Kondrashov, A. S., Sunyaev, S. & Kondrashov, F. A. (2002). Dobzhansky-Muller incompatibilities in protein evolution. *Proc Natl Acad Sci U S A* 99, 14878-83.

Mehrotra, K., Mohan, C. K., Ranka, S. & NetLibrary Inc. (1997). *Elements of artificial neural networks*. Complex adaptive systems., MIT Press, Cambridge, Mass.

Ng, P. C. & Henikoff, S. (2001). Predicting deleterious amino acid substitutions. *Genome Res* 11, 863-74.

Nolan, P. M., Peters, J., Strivens, M., Rogers, D., Hagan, J., Spurr, N., Gray, I. C., Vizor, L., Brooker, D., Whitehill, E., Washbourne, R., Hough, T., Greenaway, S., Hewitt, M., Liu, X., McCormack, S., Pickford, K., Selley, R., Wells, C., Tymowska-Lalanne, Z., Roby, P., Glenister, P., Thornton, C., Thaung, C., Stevenson, J. A., Arkell, R., Mburu, P., Hardisty, R., Kiernan, A., Erven, A., Steel, K. P., Voegeling, S., Guenet, J. L., Nickols, C., Sadri, R., Nasse, M., Isaacs, A., Davies, K., Browne, M., Fisher, E. M., Martin, J., Rastan, S., Brown, S. D. & Hunter, J. (2000). A systematic, genome-wide, phenotype-driven mutagenesis programme for gene function studies in the mouse. *Nat Genet* 25, 440-3.

Rossant, J. & McKerlie, C. (2001). Mouse-based phenogenomics for modelling human disease. *Trends Mol Med* 7, 502-7.

Rumelhart, D. E., McClelland, J. L. & University of California San Diego. PDP Research Group. (1986). *Parallel distributed processing : explorations in the microstructure of cognition*, MIT Press, Cambridge, Mass.

Russell, R. B. & Barton, G. J. (1994). Structural features can be unconserved in proteins with similar folds. An analysis of side-chain to side-chain contacts secondary structure and accessibility. *J Mol Biol* 244, 332-50.

Santibañez-Koref, M. F., Gangeswaran, R., Santibanez-Koref, I. P., Shanahan, N. & Hancock, J. M. (2003). A phylogenetic approach to assessing the significance of missense mutations in disease genes. *Hum Mutat* 22, 51-8.

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.

Wang, Z. & Moult, J. (2001). SNPs, protein structure, and disease. *Hum Mutat* 17, 263-70.

Waterston, R. H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J. F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., An, P., Antonarakis, S. E., Attwood, J., Baertsch, R., Bailey, J., Barlow, K., Beck, S., Berry, E., Birren, B., Bloom, T., Bork, P., Botcherby, M., Bray, N., Brent, M. R., Brown, D. G., Brown, S. D., Bult, C., Burton, J., Butler, J., Campbell, R. D., Carninci, P., Cawley, S., Chiaromonte, F., Chinwalla, A. T., Church, D. M., Clamp, M., Clee, C., Collins, F. S., Cook, L. L., Copley, R. R., Coulson, A., Couronne, O., Cuff, J., Curwen, V., Cutts, T., Daly, M., David, R., Davies, J., Delehaunty, K. D., Deri, J., Dermitzakis, E. T., Dewey, C., Dickens, N. J., Diekhans, M., Dodge, S., Dubchak, I., Dunn, D. M., Eddy, S. R., Elnitski, L., Emes, R. D., Eswara, P., Eyras, E., Felsenfeld, A., Fewell, G. A., Flicek, P., Foley, K., Frankel, W. N., Fulton, L. A., Fulton, R. S., Furey, T. S., Gage, D., Gibbs, R. A., Glusman, G., Gnerre, S., Goldman, N., Goodstadt, L., Grafham, D., Graves, T. A., Green, E. D., Gregory, S., Guigo, R., Guyer, M., Hardison, R. C., Haussler, D., Hayashizaki, Y., Hillier, L. W., Hinrichs, A., Hlavina, W., Holzer, T., Hsu, F., Hua, A., Hubbard, T., Hunt, A., Jackson, I., Jaffe, D. B., Johnson, L. S., Jones, M., Jones, T. A., Joy, A., Kamal, M., Karlsson, E. K., et al. (2002). Initial sequencing and comparative analysis of the mouse genome. *Nature* 420, 520-62.

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