Selective Constraints and Human Disease Genes: Evolutionary and Bioinformatics Approaches

Advanced article

Article Contents

Online posting date: 15th July 2008

Hernán Dopazo, *Centro de Investigación Príncipe Felipe, Valencia, Spain*

Natural selection rejects with variable strength, mutations reducing the individual's capability to survive and reproduce. Evolutionary theory predicts that mutations producing disease will be under strong selective constraints. Selective strength at the codon level will determine if mutation frequency will increase, decrease or change randomly during evolution. This strength finally serves in the prediction of nonsynonymous single nucleotide polymorphisms (nsSNPs) producing disease in humans. By using comparative genomics data and maximum likelihood phylogenetics approaches we demonstrate that mutations on residues showing low rates of evolution are significantly associated to disease and not to human genetic polymorphisms.

Introduction

Since the earlier works of JBS Haldane on sickle-cell anaemia, biologists recognize the power of natural selection on genetic variation and its association to human diseases. Further developments demonstrated that most of the genetic changes occurring in a population do not affect the phenotype, or more accurately, the reproductive capacity (fitness) of the genotypes carrying genetic variants (Kimura, 1983). Recently, 3.1 million single nucleotide polymorphisms (SNPs) were found in the human genome (IHMC, 2007) and a major goal on biomedical research is to understand the role of the common genetic variants in susceptibility to common diseases in human populations. **See also**: An Evolutionary Framework for Common Disease; Single Nucleotide Polymorphism (SNP)

A worldwide survey on the genetic variation in genes associated to common human diseases concluded that SNPs occur at a frequency of 1 out of 346 bp and are roughly equally divided between synonymous and nonsynonymous changes (Cargill *et al*., 1999). As approximately, two-thirds of random mutations in coding sequences alter an amino acid, the fact that nsSNPs compromise one half the total SNPs, implies strong selection against amino acid altering changes. The force of selection is also evident when comparing nsSNPs causing nonconservative amino acid substitutions with those causing a conservative change. Nonconservative nsSNPs represent only 36% of all nsSNPs, whereas randomly distributed mutations would be expected to produce a higher proportion (52%) of nonconservative changes (Cargill *et al*., 1999). Currently, the NCBI SNP database (dbSNP, built 127) collects 5 689 286 validated human SNPs out of which 78 845 are nonsynonymous coding SNPs (nsSNPs). That means that about only 1% of the human validated SNPs could probably affect gene function. One of the most important questions in human genetics is to deduce which of these genetic variants are functionally relevant for human health. In other words, which of these 1% of genetic variants are targets of selective or neutral evolutionary processes in the human genome. Far from the theoretical interest of this enquire, this prediction would help in the genotyping process of SNPs probably associated to disease in classical genotype–phenotype association studies in human populations. **See also**: Evolution: Neutralist View; Mutations in Human Genetic Disease

In this article, I will overview many of the main methods developed to predict the functional properties of nsSNPs in the human genome. I will insist that most of these methods are based on features derived from protein structures and/or the evolutionary conservation which represent proxies to infer its cost on fitness. Next, I will focus on the measure and description of the selective constraints associated to nsSNPs in the human genome as a direct way to search for

ELS subject area: Evolution and Diversity of Life

How to cite:
Dopazo, Hernán (July 2008) Selective Constraints and Human Disease Genes: Evolutionary and Bioinformatics Approaches. In: Encyclopedia of Life Sciences (ELS). John Wiley & Sons, Ltd: Chichester.
DOI: 10.1002/9780470015902.a0020762

selective pressures acting at the codon level. We will see that this evolutionary approach based on the maximum likelihood (ML) estimation of evolutionary rates evaluates the fitness of each nsSNPs at the codon level at the same time that improves previous attempts to differentiate deleterious alleles from neutral polymorphisms in the human genome. Finally, I will present the PupaSNPs suite tool where all the predictions for nsSNPs on coding sequence are collected for the human genome. **See also**: Amino Acid Substitutions: Effects on Protein Stability; Molecular Evolution: Rates

Functional Prediction of nsSNPs

The earliest studies on the functional prediction of human nsSNPs were pioneered by Sunyaev *et al*. (2000, 2001), Chasman and Adams (2001), Wang and Moult (2001), Miller and Kumar (2001), Saunders and Baker (2002) and Santibáñez Koref *et al*. (2003). A rough description of the main publications is shown in **Table 1**.

Sunyaev *et al*. (2000) estimated that approximately 70% of disease-causing mutations occur at structurally and functionally important sites with well-defined properties such as less than 5% of solvent accessibility, sometimes located in β strands, active sites, disulfide bonds or evolutionary conserved sites. Moreover, they found that most of the allelic variants map to the same structurally and functionally important regions of the proteins suggesting that many of them probably have negative effects on the phenotype. In a subsequent study, Sunyaev *et al*. (2001) estimated that approximately 20% of the human nsSNPs affect protein function and that an average human genotype carries about 2000 of such nsSNPs. They observed that the majority of disease-causing mutations were at low frequencies in human populations (1–20%), which was considered as a validation of their method.

Wang and Moult (2001) found that by far the largest proportion (83%) of disease-nsSNPs affects protein stability, 5% maps on binding sites and approximately 10% correspond to cases where their three-dimensional (3D) structural model gives a false-negative result. They predicted that 70% of nsSNPs studied in hypertension, cardiovascular, endocrinology and neuropsychiatric diseases correspond to cases of neutral polymorphisms whereas the remaining 30% affect the stability of the protein. Alternatively, Chasman and Adams (2001) used a combination of statistical methods to define structural and evolutionary parameters with significant association to disease. From the knowledge of the effects of about 6000 mutations from the Lac repressor and the T4 lysozyme protein they estimated that approximately 26–32% of nsSNPs have deleterious effects on human protein function.

Although most of these studies mainly focused on structural parameters of proteins, Miller and Kumar (2001) explicitly studied the role of the evolutionary conservation in the functional prediction of nsSNPs emphasizing the risk of using concepts like conservation profile and similarity cutoff percentage values used in the previous models. They pointed out that evolutionary data cannot be treated as independent observations for use in statistics because they share a nonrandom structure of dependence defined in the historical relationships of the species (Felsenstein, 1985). That is, model approaches based on similarity could overestimate the variability of a given site if an identical residue appears in multiple species due to phylogenetic constraints. Moreover, the alignment profile score could underestimate the amount of variation in the sequences using highly conservative cutoff values. Therefore, using an explicit method of phylogenetic reconstruction on seven human disease proteins, Miller and Kumar (2001) demonstrated that human nsSNPs mutations are overabundant at amino acid positions most conserved throughout the long-term history of metazoans. Human polymorphic replacement mutations and silent mutations were found randomly distributed across sites with respect to the level of conservation of amino acid sites within genes. They concluded that disease-causing amino acid changes are those that are not observed among species probably because they are not accepted by natural selection in long-term evolutionary time.

In the same vein, an explicit statistical phylogenetic model was developed by Santibáñez Koref *et al*. (2003). The method indicates the probability of a given mutation being pathological considering the evolutionary conservation and the variability associated to each protein. Although the method they developed was outstanding in the fields of comparative genomic and human health, the necessary calibration of the model for each human protein is a major drawback for its use in large-scale analysis. **See also**: Comparative Genomics

Saunders and Baker (2002) evaluated the behaviour of alternative variables in the functional prediction of nsSNPs. When using a combination of evolutionary and structural variables, they concluded that the prediction is better than when a single kind of variable is used on its own. When fewer than 5–10 homologues are available, they emphasized that the prediction of deleterious mutation should include structural information, suggesting that the evolutionary data is more informative than the structural data when a high number of sequences are used for prediction. **See also**: Homology in Character Evolution

Finally, Arbiza *et al*. (2006) introduced an explicit evolutionary measure of selective pressures at a codon level as a direct functional predictor of nsSNPs. Using ML models to estimate the well-established ratio (ω) between nonsynonymous to synonymous rates of evolution in mammals, they concluded that codons with $\omega < 0.1$ maps residue where mutations producing disease are frequent in human. **See also**: Synonymous and Nonsynonymous Rates

Methods and Web Resources

Most of the methods used for the functional prediction of nsSNPs are characterized by the use of the structural (ST), sequence (SQ) and/or functional (FN) information of

Table 1 Studies analysing the main variables associated to the functional prediction of nsSNPs

Publication	Parameters and statistics	Disease mutations	Control/prediction dataset	Main conclusions
Sunyaev *et al*. (2000)	Structural features such as solvent accessibility, secondary structure, active sites and disulfide bonds were combined with conservation features taken from homologous sequences. χ^2 statistical analyses compare nsSNP distributions	Disease dataset: 551 nsSNPs taken from Swiss-Prot, OMIM and PDB. Allelic variation dataset: 86 nsSNPs taken from Swiss-Prot, OMIM, HGBASE, Chakravarti dataset and PDB	Homologues dataset 1: 225 nsSNPs taken from close relatives of disease genes. Homologues dataset 2: 261 nsSNPs taken from close relatives of human allelic variation dataset genes	Disease-causing mutations often affect intrinsic structural features of proteins. Approximately 70% of the disease-causing mutations are located in sites likely to be structurally and functionally important. The fraction of polymorphic sites located in structurally and functionally important regions was 45%, which is significantly higher than the 24% in the case of the interspecies variation. Allele frequency distribution suggests that variants in structurally important sites are not selectively neutral
Chasman and Adams (2001)	Sixteen features contribute to the model. Among the continuous variables are those such as residue accessibility, relative residue entropy and relative residue B-factor. Among the categorical factors such as unusual AA, unusual AA by class and rare AA. Turn or helix breaking. Buried residues, conserved position, etc. Phylogenetic entropy deduced from HSSP files. Statistical analyses use ANOVA *F*-statistic for continuous and χ^2 test for categorical variables. Probabilistic models with combination of variables was used to predict functional characteristics of nsSNPs	Lac repressor: $\sim$4000 nsSNPs T4 lysozyme: $\sim$2000 nsSNPs	SNPs survey from Case Western Reserve University & Whitehead cSNP databases	The variables used in the study are strong predictors of an effect on function for lac repressor and lysozyme. They estimate that approximately 26–32% of nsSNPs probably affect the function of human proteins

(*Continued*)

Table 1 Continued

Publication	Parameters and statistics	Disease mutations	Control/prediction dataset	Main conclusions
Wang and Moult (2001)	3D model structures of proteins were built using comparative structural modelling. Defined rules for assessing the effects of nsSNPs based on protein stability, ligand binding, catalytic position, allosteric regulation and post-translational modifications	SNPs disease set: 23 proteins, 262 nsSNPs, derived from NIH dbSNP, HMDB, PDB	SNP population set: 22 proteins, 42 nsSNPs, derived from Case Hypertension candidate genes and the Whitehead cSNP databases (cardiovascular, endrocrinology and neuropsychiatric diseases), PDB	83% of disease mutations affect some of the 12 variables associated to protein stability. According to the features used in the model, 70% of the SNP population set has no effect on protein function being selectively neutral
Miller and Kumar (2001)	Evolutionary analysis of nsSNPs associated to human disease. Phylogenetic tree reconstruction of eukaryote genes and amino acid relative frequencies change	Genetic variation (1004 nsSNPs) producing disease from seven disease human genes. Disease (10 262)-associated mutations obtained from HGMD	Polymorphic (50 nsSNPs) and silent variation (94 nsSNPs) taken from the same seven disease-associated genes	Human replacement mutations resulting in disease are overabundant at amino acid positions most conserved throughout the long-term history of metazoans. Human polymorphic replacement mutations and silent mutations are randomly distributed across sites with respect to the level of conservation of amino acid sites within genes. Disease-causing amino acid changes are of types usually not observed among species
Saunders and Baker (2002)	Structural and evolutionary features selected to test their relevance in the nsSNP prediction problem. Among them, percentage of solvent accessibility area, normalized B-factor, residue burial, Sunyaev structural rules, Blosum62 relative frequencies, normalized size entropy and SIFT predictions	Deleterious nsSNPs mutations from Lac repressor: 1166, T4 lysozyme: 175 and HIV-1: 159; adding 1500. 191 disease alleles taken from OMIM and Swiss-Prot40	Neutral nsSNPs mutations from Lac repressor: 2255, T4 lysozyme: 1340 and HIV-1: 111; adding 3706. 87 neutral alleles taken from OMIM and Swiss-Prot40	Methods for deleterious mutation prediction should include structural information when fewer than 5–10 homologues are available. *Ab initio* predicted structures may be useful in such cases when high-resolution structures are unavailable and few homologous sequences exist

Santibanez Koref *et al*. (2003)	The method describes a formal statistical framework to assess nsSNPs using phylogenetic analysis of mammal genes, codon sequences and physicochemical amino acid properties	The TP53 mutation database 1038 alleles containing nsSNPs	1047 mutations associated to interspecies variation	A *Z*-score indicates the probability that a given mutation is pathological considering evolutionary conservation and variability
Arbiza *et al*. (2006)	Evolutionary analysis of nsSNPs associated to human disease. Codon substitution model, Ensembl-Database orthologous relationships. Site-specific maximum-likelihood models ($\omega = dN/dS$) run in the codeml program from PAML. Mammal and vertebrate trees	The TP53 mutation database containing 18 145 mutations	43 genes summing up to 8970 mutations derived from IDR, MeCP2 and COSMIC databases	The Kolmogorov–Smirnov test differentiates selective pressures where mutations associated to disease are more frequent than mutations not associated to disease. Selective pressures on nsSNPs with $\omega < 0.1$ are statistically associated to disease

proteins. All of them use homologous sequences (many of them without a formally differentiating orthology and paralogy) since the computation of the conservation properties or scores requires a comparison with other related proteins commonly found by blast search family methods.

Table 2 shows 12 of the most popular bioinformatic tools for the functional prediction of nsSNPs. These tools use alternative classification methods to decide which of the nsSNPs may have deleterious or neutral phenotypes. As we will see they make use of various approaches including cutoff values (SIFT, PANTHER, SNPeffect), decision trees (POLYPHEN, LS-SNP) or machine learning methods such as neural networks (NNs) (PMUT, SNAP), support vector machines (SVMs) (SNP3D, PhD-SNP, SAP) and random forests (nsSNP Analyzer). **See also**: Neural Networks; Phylogenetic Footprinting; Proteins: Mutational Effects in

Cutoff value-based methods

The *SIFT* (sorting intolerant from tolerant, http://blocks.fhcrc.org/sift; Ng and Henikoff, 2001, 2003) algorithm takes a query sequence and uses multiple alignment information to predict tolerant and deleterious substitutions for every position of the query sequence. Once the multiple alignment of sequences is done SIFT computes the PSSM (position-specific substitution matrix) containing individual probabilities for each amino acid of the protein being changed for any other of the 20 amino acids. Based on the frequencies with which the changes are observed in the alignment, *SIFT* provides the normalized probabilities for all the possible substitutions at each position of the alignment. Substitutions at each position with normalized probabilities less than a chosen cutoff are predicted to be deleterious and those greater or equal to the cutoff value are predicted to be tolerant.

Panther (http://www.pantherdb.org/tools/csnpScoreForm.jsp; Thomas and Kejariwalet, 2004) is a library of protein families and subfamilies derived by the use of Hidden Markov Model (HMM) techniques indexed by a vocabulary of more than 500 biological functional terms. The Panther library contains predictions of the effects of nsSNPs on protein functions. Panther (Thomas *et al*., 2003) uses the alignment and tree of each family and subfamily to compute the position-specific evolutionary conservation (PSEC) score. This score works as a 'functional likelihood' value of amino acid substitution in a protein family. The authors calibrated the scores and concluded that alleles with a cutoff value smaller than -3 (subPSEC < -3) correspond to deleterious mutations.

SNPeffect (http://snpeffect.vib.be/; Reumers *et al*., 2006) is a database containing predictions from sequence and structure-based bioinformatic tools of nsSNPs. SNPeffect analyses the effect of SNPs on three categories of functional properties: (1) structural and thermodynamic properties affecting protein dynamics and stability, (2) the integrity of functional binding sites and (3) changes in post-translational processing and cellular localization of proteins. The bioinformatic tools and databases used for such predictions are FOLDX, TANGO, AmyScan, PROF, Hsp70, CSA, Phosphobase, O-GlycoBase, *N*-terminal rule and PA Subcellular (see references in Reumers *et al*., 2006). SNPeffect provides alternative ranges of values being considered disease-associated or neutral.

Decision tree-based methods

Polyphen (polymorphism phenotyping, http://coot.embl.de/PolyPhen/; Ramensky *et al*., 2002) is a sequence and structural based algorithm for the functional prediction of nsSNPs. First, Polyphen characterizes the substitution sites at a structural level by looking for information in the human section of the SWALL database (a comprehensive protein sequence database that combines the high quality of annotation in Swiss-Prot and all the protein-coding sequences from the EMBL nucleotide sequence database). Second, Polyphen computes the PSIC (*position-specific independent counts*) matrix scores. Elements of the matrix (profile scores) are logarithmic ratios of the likelihood of finding a given amino acid at a particular position to the likelihood of finding this amino acid at any position (background frequency). Polyphen computes the absolute value of the difference between profile scores of both allelic variants in the polymorphic position. Big values for this difference may indicate that the studied substitution is rarely or never observed in the protein family. Third, mapping amino acid substitutions on proteins with recognized tertiary and quaternary structure, the algorithm searches for structural changes and contacts sites with other proteins. Finally, with all the information, Polyphen uses empirically derived rules to predict that an nsSNP is probably damaging, possibly damaging, benign or unknown.

LS-SNP (large-scale human SNP annotation, http://alto.compbio.ucsf.edu/LS-SNP/; Karchin *et al*., 2005) is a database collecting results from a pipeline that maps nsSNPs onto protein sequences, functional pathways and comparative protein structure models, and predicts positions where nsSNPs destabilize proteins, interfere with the formation of domain–domain interfaces, have an effect on protein–ligand binding or severely impact human health. By integrating information based on sequence, evolution and structure with a combination of knowledge-based rules and an SVM, LS-SNP predicts positions where amino acid substitutions destabilize protein structure.

Machine learning based methods

PMUT (http://mmb2.pcb.ub.es:8080/PMut/; Ferrer-Costa *et al*., 2005) performs its predictions by retrieving a series of structural parameters such as volume parameters, secondary structure propensities, hydrophobicity descriptors and sequence potential, among others. Comparative descriptors come from the scoring mutation matrices (PAM40 and BLOSUM62) and from the multiple sequence alignment found by using two iterations of PSI-Blast running over a

Table 2 Comparison of some of the methods used in the functional prediction of nsSNPs

Methods	Algorithm features	Substitution score	Homology	Classifcation	nsSNPs functional prediction	
SIFT http://blocks.fhcrc.org/sift	SQ	Multiple alignment of homologous sequences and posterior computation of probabilities for each substitution site	PSSM	PSI-BLAST Swiss-Prot and TrEMBL	Normalized cutoff value	Disease association based on normalized cutoff probabilities ($p < 0.05$)
POLYPHEN http://coot.embl.de/PolyPhen/	ST	Secondary structure, solvent accessible area, φ-ψ dihedral angles and contact sites inference among other parameters	PSIC	BLAST NRDB	Decisional tree	Categorical: benign, probably damaging or possibly damaging
PANTHER http://www.pantherdb.org/tools/csnpScoreForm.jsp	FN	Panther functional annotation. HMMs based. Family tree	subPSEC	Family–subfamily HMMs definition	subPSEC cutoff value	Deleterious if subPSEC < -3. More negative values predict more deleterious substitutions
SNPs3D http://www.snps3d.org	ST	A total of 15 stability factors (continuous and binaries), such as cavity formation, loss of disulfide bridge, crystallographic temperature and others contributing to energy and entropy	PSSM	PSI-BLAST Swiss-Prot	Two SVMs (structure stability and sequence profile)	A negative SVM score indicates deleterious mutations. Accuracy is significantly higher when both SVMs agree
LS-SNP http://alto.compbio.ucsf.edu/LS-SNP	ST	Structural features from structural protein modelling (such as solvent accessibility, buried charge and *in silico* mutation-violated spatial restraints) and amino acid residues (change in residue	Relative entropy values derived from HMM	PSI-BLAST Swiss-Prot and TrEMBL and nrNCBI	Decisional tree and SVM	A negative SVM score indicates deleterious mutations

(*Continued*)

Table 2 Continued

Methods	Algorithm features	Substitution score	Homology	Classifcation	nsSNPs functional prediction	
		volume, charge, hydrophobicity and Grantham values)				
PMUT http://mmb2/pcb.ub.es:8080/PMut	ST	Secondary structure location, solvent accessibility, residue size, free energy of water to octanol transfer and secondary structure propensity	Shannon entropy, average mutation matrix score and PSSM (Bolsum62)	PFAM PSI-BLAST nrSwiss-Prot and TrEMBL	Neural network	Pathological index ranges from 0 to 1 (indexes > 0.5 signal pathological mutations). Additionally, a confidence index ranges from 0 to 9
PhD-SNP http://gpcr2.biocomp.unibo.it/cgi/predictors/PhD-SNP/PhD-SNP.cgi	SQ	Sequence information taken from a centred window of 19 residues from the nsSNP	Transition frequencies of wild-type and mutant residues	BLAST NRDB	2 SVMs (profile and sequence) in the Hybrid method	Hybrid SVM predicts deleterious or neutral mutations according to a reliability index (RI)
SNAP http://cubic.bioc.columbia.edu/services/SNAP	ST SQ	Solvent accessibility, chain flexibility, sequence information taken from a window centred in the nsSNP, transition frequencies of wild-type and mutant triplets, Swiss-Prot annotations	PSIC-Blosum62	PSI-BLAST PFAM	Neural network	NN predicts neutral or nonneutral mutations according to an RI and the associated expected accuracy
nsSNP Analyzer http://snpanalyzer.utmem.edu/	ST	Solvent accessibility, environmental polarity and secondary structure	PSSM + SIFT predictions	Structural ASTRAL	Random Forest	Disease or neutrality are predicted in association with the SIFT cutoff probability value
SAP http://sapred.cbi.pku.edu.cn	ST	Solvent accessibility, difference between wild-type and mutant structural (3D) neighbour profiles, nearby functional sites, energy model, number of hydrogen bonds, disulfide	A slight variation of SIFT scores and conservation score	Swiss-Prot	SVM	Two types of SVM predictions. One is based on both the structural and sequence information, the other relies on the sequence information only

		bonds, disordered region, aggregation properties, HLA family				
SNPeffect http://snpeffect.vib.be	ST	Energetic effects of nsSNPs, changes in protein aggregation or amyloidosis are evaluated using FoldX force field, TANGO and AnyScan. Active sites are located by means of the Catalytic Site Atlas database. Subcellular localization predicted by PA Subcellular and Psort II. Post-translational modification using PhosphoBase, O-Glycobase and other databases	Zvelebil truth table for amino acid properties	BLAST Swiss-Prot	Cutoff values	nsSNPs are characterized as neutral or deleterious according to different cutoff values associated to structure and dynamics, functional sites and cellular processing
SeqProfCod http://sgu.bioinfo.cipf.es/services/Omidios/	SQ	Sequence information taken from a centred window of 19 residues from the nsSNP. Site-specific ML models ($\omega = dN/dS$). Mammals phylogenetic tree information	Codon substitution model and residues transition frequencies	Ensembl-Database orthologues	Sequence-Profile-Codon model	Sequence-Profile-Codon SVM predicts deleterious or neutral mutations according to RI values

nonredundant Swiss-Prot/trEMBL database. PMUT implemented two NNs as predictor engines. Both NNs were trained with human mutational data. Irrespective of the NN, the final output is (1) a pathogenicity index ranging from 0 to 1 (mutations associated with an index above 0.5 are taken as pathological) and (2) an index ranging from 0 (low) to 9 (high) corresponding to the confidence level of the prediction.

SNAP (screening for nonacceptable polymorphism, http://www.rostlab.org/services/SNAP/; Bromberg and Rost, 2007) is an NN-based method that uses a variety of biophysical characteristics associated to the substitutions, as well as evolutionary information, to make functional predictions regarding mutated proteins. The network takes protein sequences and lists of mutants as input, returning a score for each substitution. These scores are translated into binary predictions of effect (neutral/nonneutral), reliability indices (RIs) and expected accuracy. RIs are indicative of confidence in prediction, whereas the expected accuracy is a number of correctly predicted (at a given RI) neutral or nonneutral samples in the SNAP testing set.

SNPs3D (http://www.snps3d.org/; Yue *et al*., 2006) is a server tool which assigns molecular functional effects of nsSNPs based on structure and sequence analysis using two SVM models. The first model (the stability model) is based on the hypothesis that many disease SNPs affect protein function primarily by decreasing protein stability. The second model (the profile model) is based on the analysis of homology sequence families related to human proteins. SNPs3D database contains the prediction of nsSNPs of the NCBI dbSNP database. They predicted that approximately 30% of these mutations are associated to human diseases.

PhD-SNP (predictor of human deleterious SNP, http://gpcr.biocomp.unibo.it/~emidio/PhD-SNP/PhD-SNP.htm; Capriotti *et al*., 2005) is a sequence-based method using two different SVM classifiers. The SVM-Sequence based method use sequence information taken from a centred window of 19 residues from the nsSNPs. This SVM is coupled to SVM-Profile trained on sequence profile information in the Hybrid method. The SVM-Sequence and the Hybrid methods predict deleterious or neutral nsSNPs according to an RI. Although PhD-SNP does not make any inference from structural parameters, it seems to outperform other cutoff-based value predictors based on sequence, function or structure such as SIFT or PANTHER.

SAP (single amino acid polymorphism, http://sapred.cbi.pku.edu.cn/; Ye *et al*., 2007) is an SVM tool which is characterized by using, aside from the well-recognized prediction power of variables such as sequence conservation and solvent accessibility, new biologically informative attributes including structural neighbour profiles, nearby functional sites and aggregation properties among others. The new attributes studied by SAP provide insights into the mechanisms of the disease association of nsSNPs. SAPRED web server requires two PDB format files describing the structures of the wild-type and variant proteins. For proteins with no structural information, an alternative method called SAPRED_SEQ makes the prediction based on sequence-derived attributes only.

nsSNPAnalyzer (http://snpanalyzer.utmem.edu/; Bao *et al*., 2005) uses a machine learning method called Random Forest (Breiman, 2001) to classify nsSNPs as deleterious or neutral. It uses information from the structural environment of the SNP, the normalized probability of the substitution in the multiple sequence alignment and the similarity between the original amino acid and mutated amino acid. nsSNPAnalyzer searches for homologous protein structures since it does not work without structural information.

Natural Selection and Disease

As we have described earlier, all these methods try to predict the functional consequences of nonsynonymous mutations occurring in a protein by means of the use of different sequence and/or structural parameters. This information is finally used as a proxy for the definition of selective constraints posed by natural selection on the protein site where the nonsynonymous mutations occurred.

Natural selection shapes the genetic variation of the population according to the functional role played by the new mutant that will finally be accepted or discarded from the gene pool. A common approach to determine the selective pressures acting at a molecular level is the estimation of the ratio of nonsynonymous to synonymous rates of substitution ($\omega = dN/dS$). An estimation of dN that is significantly different from that of dS, provides convincing evidence of a nonneutral evolutionary process. Codon-based ML models allow the study of natural selection in a site-by-site approach thus providing estimates of selection at a codon level (Yang, 2003). **See also**: Synonymous and Nonsynonymous Rates

Here we will show that the estimation of the selective pressures on a well-known protein model (p53) associated to human disease (cancer) can be computed at a codon level using well-known statistics methods of evolutionary biology. These constraints, having worked throughout millions of years on orthologous sequences, can be successfully used as a predictor of the phenotypic effects of cSNPs. Secondly, we will demonstrate that nsSNPs with selective strength (ω) smaller than 0.1 are frequently associated to human disease (Arbiza *et al*., 2006). Finally, in a pure bioinformatics framework, we will see that by using a trained machine learning algorithm applied on sequence data increases the likeliness of association to disease (Capriotti *et al*., 2008). This information mapped on the database of human SNPs provides us with a full prediction of all the coding variation producing disease in the human genome. This information is currently available at the PupaSuite server: http://pupasuite.bioinfo.cipf.es (Conde *et al*., 2006).

Mutations and constraints in p53

The p53 tumour suppressor is a 393-amino acid transcription factor that activates the transcription of a number of downstream genes. Structurally and functionally, it can be divided into five regions: an acidic *N*-terminal transactivation domain (p53TA, residues 1–60), a proline-rich domain (p53PR, residues 61–97), a hydrophobic DNA (deoxyribonucleic acid)-binding domain (p53DB, 100–300), a tetramerization domain (p53TR, 320–360) and a basic *C*-terminal domain (p53CO, 361–393). The IARC TP53 mutation database collects the largest number of mutations in this protein. **Figure 1** shows that codon mutation are scattered throughout the coding sequence, although 96% of them (17 389/18 135) cluster within the p53DB domain. Six different 'mutational hotspots' (defined by a mutation frequency higher than 2% of all mutations) have been identified at residues Arg^{175}, Gly^{245}, Arg^{248}, Arg^{249}, Arg^{273} and Arg^{282}. According to this description, these mutational hotspots fall within the p53DB domain, and since they are structurally relevant to protein function (Cho *et al.*, 1994), they would be expected to be protected against mutations by strong purifying selection (Golding, 1994). **See also**: Tumour Suppressor Genes

ML adjustment of evolutionary parameters using the M8($\beta+\omega$) (hereafter M8) selection model from CodeML program in PAML (Yang, 2007) suggests that almost 100% of p53 codon sites are constrained under the influence of purifying selection, and only a minimum proportion of sites evolved with $\omega>1$. The parameters of the β distribution suggest that ω describes an 'L-shaped' curve over sites, with most sites in p53 being highly conserved. Posterior probabilities obtained from the empirical Bayes approach were not significant ($p<95\%$) for any of the protein residues, suggesting the absence of positive selection sites (PSS) on p53 protein. When the alternative sitewise likelihood-ratio method (SLR, Massingham and Goldman, 2005) is used to fit the selective constraints at the codon level, the program found 228 codons sites under the influence of strong purifying selection after correcting for multiple testing (202 at $p<0.01$ and 26 at $p<0.05$) (see Arbiza *et al.*, 2006 for a full description of the methods). It is interesting to notice that this number is higher than the 109 (47.80%) that are, actually, phylogenetically conserved (which are those generally used in the methods described previously) and never change in amino acid identity during evolution. **Table 3** summarizes data related to p53 domains, codons, indels, mutations and statistics. Independent of the model used to estimate ω, p53DB and p53TR domains showed the highest number of cancer mutations which were associated with the lowest median and mean ω values observed in the analysis. One-tail Kolmogorov–Smirnov (K–S) tests demonstrated that the p53DB and p53TR domains have a significantly low ω value distribution ($p<0.05$) in comparison with the rest of the p53 domains. Although the mean estimation was close to 0.1 in both domains, the distribution of ω values in the p53DB was lower than in the p53TR domain (data not shown).

In summary, estimates of natural selection acting on p53 coding sites statistically differentiate the relevant functional domains where the prevalence of cancer mutations are the highest in a protein. This pattern is what would be expected if relevant functional structures had been constrained, during evolution, under strong selective forces avoiding nonsynonymous changes.

Selective constraints in p53 structure

More than 40 years ago, Zuckerkandl and Pauling (1965) proposed that a protein sequence will evolve at a rate primarily determined by the proportion of sites involved

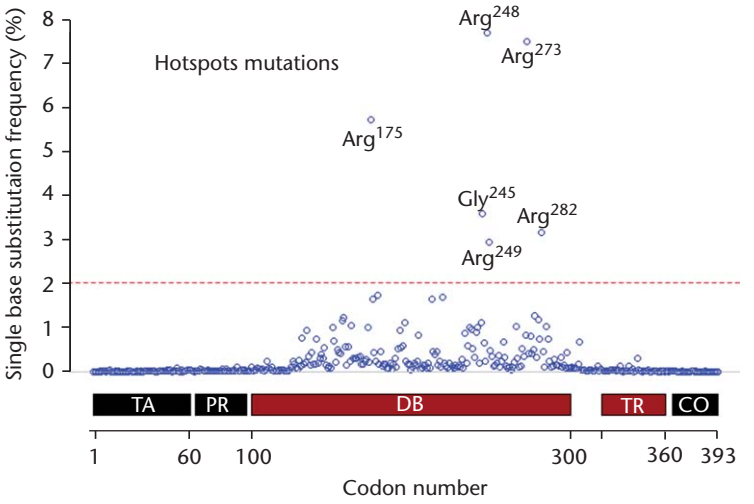

Figure 1 Distribution of p53 mutations. Mutation frequencies collected in the IARC TP53 R10 database (18 145 nonsynonymous mutations) are plotted against the protein domains. The DNA-binding (p53DB) domain contains six residues considered mutational hotspots in cancer. Reproduced from Arbiza *et al.* (2006), Copyright Elsevier (2006).

Table 3 Summary of p53 domains, mutations and statistics according to M8 model and SLR method. Mutations were deduced from the IARC TP53 database

Alignment			Mutations			ω Statistics			
Domain	Codon	Indels	Total	Mps[a]	Model	Minimum	Median	Mean	Maximum
TA	1–60	38	96	1.6	M8	0.030	0.334	0.379	1.747
					SLR	0.000	0.269	0.369	1.865
PR	61–97	22	151	4.2	M8	0.029	0.314	0.376	1.338
					SLR	0.000	0.307	0.376	1.447
DB	100–300	5	17 389	87.0	M8	0.027	0.039	0.116	1.423
					SLR	0.000	0.029	0.095	2.018
TR	325–355	0	178	5.1	M8	0.028	0.067	0.126	0.456
					SLR	0.000	0.068	0.103	0.379
CO	361–393	11	18	1.6	M8	0.027	0.216	0.255	0.878
					SLR	0.000	0.176	0.226	0.882

[a]Mean number of mutations per site.

in specific functions. To study this well-supported theoretical prediction in p53 residues, we studied the selective constraints of the core domain in complex with DNA. By defining three ranges of ω values (red: $0 \leqslant \omega \leqslant 0.1$, orange: $0.1 < \omega \leqslant 0.2$ and yellow: $0.2 < \omega \leqslant 0.3$), we labelled residues in the structures. According to the same expectation residues where neutral or nearly neutral evolution was deduced (labelled green, $\omega > 0.30$), are expected not to form part of the relevant functional domains of the protein. **See also**: Molecular Clocks; Molecular Evolution: Nearly Neutral Theory; Molecular Evolution: Neutral Theory

Figure 2a shows the distribution of the ω_{SLR} values, on the core p53DB domain structure. **Figure 2b** depicts a schematic representation showing the primary sequence and the secondary structure where the most relevant residues are shown. For a full description of constraints on p53DB and p53TR domains readers can read Arbiza *et al.* (2006). Here we overview the most relevant conclusion emphasizing that residues where denaturizing mutants are observed (Pro^{143}, Arg^{175}, Gly^{245}, Arg^{249}, Glu^{258} and Arg^{282}) (red circles), and those involved in Zn^{2+} coordination (Cys^{176}, His^{179}, Cys^{238} and Cys^{242}) (white circles) coincided with red labels, suggesting that they are under strong evolutionary constraints imposed by purifying selection. These residues are phylogenetically conserved in the alignment (marked by the star symbol: '*'), and purifying selection (PFS) was detected on them using SLR at 99% confidence level (marked by the admiration symbol: '!'). Interestingly, residues where conservation is variable, relaxation of selective constraints was deduced. In agreement with our expectations, most of these sites are placed in the external region of the structure (**Figure 2a**), outside the β strand or helix regions (**Figure 2b**).

In summary, there is a large agreement between the functional relevance of residues deduced from the estimation of selective constraints using ML models and the functional or structural importance demonstrated experimentally in p53DB domain (Cho *et al.*, 1994). Moreover, we found no evidence that residues with neutral or nearly neutral values of $\omega > 0.30$ (green labelled) play functionally or structurally important roles in p53.

Selective constraints and mutation frequency

Evolutionary biologists maintain that natural selection works in proportion to the number of deleterious mutations occurring in the population (Kimura, 1983). Frequent mutations on residues with relevant functional biochemical roles must be targeted by purifying selection and consequently would be expected to show the highest selective constraints in the protein. Otherwise, sites changing under neutral or nearly neutral evolution will not necessarily compromise major functional roles of the protein, and consequently would rarely be expected to be found associated to disease. Therefore, the pattern in the distribution of mutational frequency against ω values should likely approach an 'L-shaped' curve. **See also**: Kimura, Motoo

Using the frequency distribution of the mutations collected at the IARC TP53 database and the ω values computed for each residue, we demonstrate that p53 residues fit the predicted pattern described earlier (**Figure 3a**). As expected, disease-associated mutational hotspots observed in **Figure 1** have shown the lowest ω values ($\omega_{SLR} = 0$, $\omega_{M8} \leqslant 0.033$) observed in the study (**Figure 3b**) as a consequence of the high evolutionary constraints imposed by natural selection. It is interesting to emphasize that there were no residues showing high ω values ($\omega > 0.3$, classically considered the minor neutral limit of selection) that also showed a high frequency of mutations associated to human disease (freq > 0.5). Finally, according to the distribution of the residues deduced under the influence of purifying selection at 95% or 99% statistical confidence (red dots in **Figure 3b**), we define a cutoff representing an a priori hypothesis to detect residues associated to human disease. This hypothesis suggests that residues showing $\omega < 0.1$ are always under the influence of the highest purifying selection process and mutations on these sites are probably

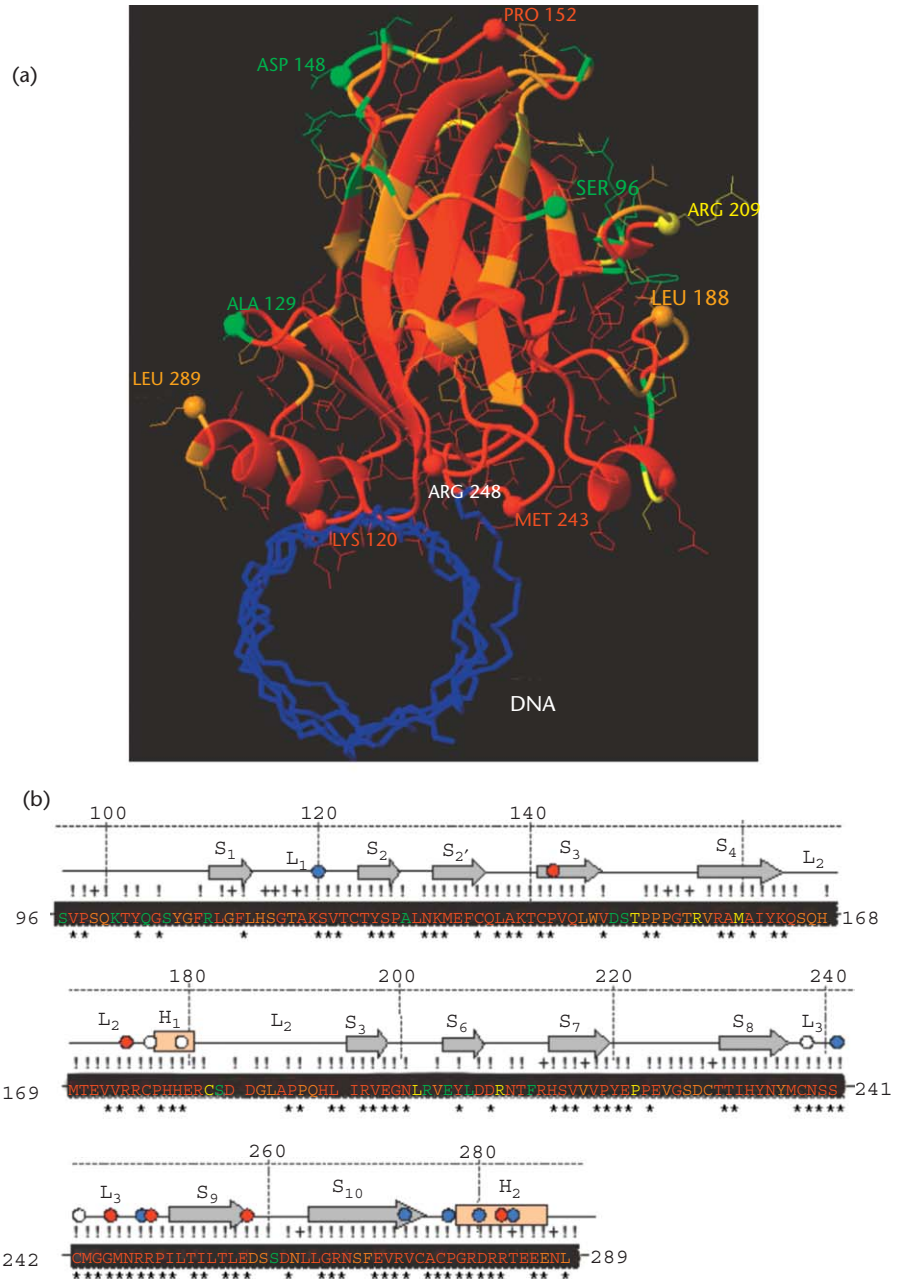

Figure 2 Mapping of selective constraints in the p53DB domain. (a) The three-dimensional structure of the p53DB domain showing residues coloured according to different selective pressures. (b) Primary amino acid sequence and secondary structure elements of the p53DB domain. Residues in red, orange, yellow and green show the gradual distribution of the selective constraints represented by ω_{SLR} values. Residues in red ($0 \leqslant \omega < 0.1$) and orange ($0.1 \leqslant \omega < 0.2$) are generally associated to DNA contact sites (blue circles), Zn^{2+} contact (white circles) and sites where mutants are known to be denaturizing (red circles) among others. A few of the sites seem to be below the limit considered for selective constraints (yellow, $0.2 \leqslant \omega < 0.3$). Residues where selective constraints were predicted to be low (green, $\omega > 0.3$) are distributed along the external regions of the core domain, and most of them are interspersed between β sheets and helices. Arg^{248} binds in the minor groove of the DNA. Ser^{185} was conserved in the cluster of primates and rodents, but was discarded in the analysis due to gap insertions in the basal species. ∗, phylogenetically conserved residue; +, SLR detected PFS at 95% confidence after correcting for multiple testing and !, as in +, but at 99% confidence. See text for a detailed explanation. Reproduced from Arbiza *et al.* (2006), Copyright Elsevier (2006).

always associated to disease. Later, we show that this hypothesis derived only using p53 protein data is a common pattern observed for amino acid mutations associated to human disease genes.

If selection has modelled ω values for generations by rejecting deleterious mutations associated to the more frequent disease mutations in the population, we would expect a gradual increase in the selective constraints in p53

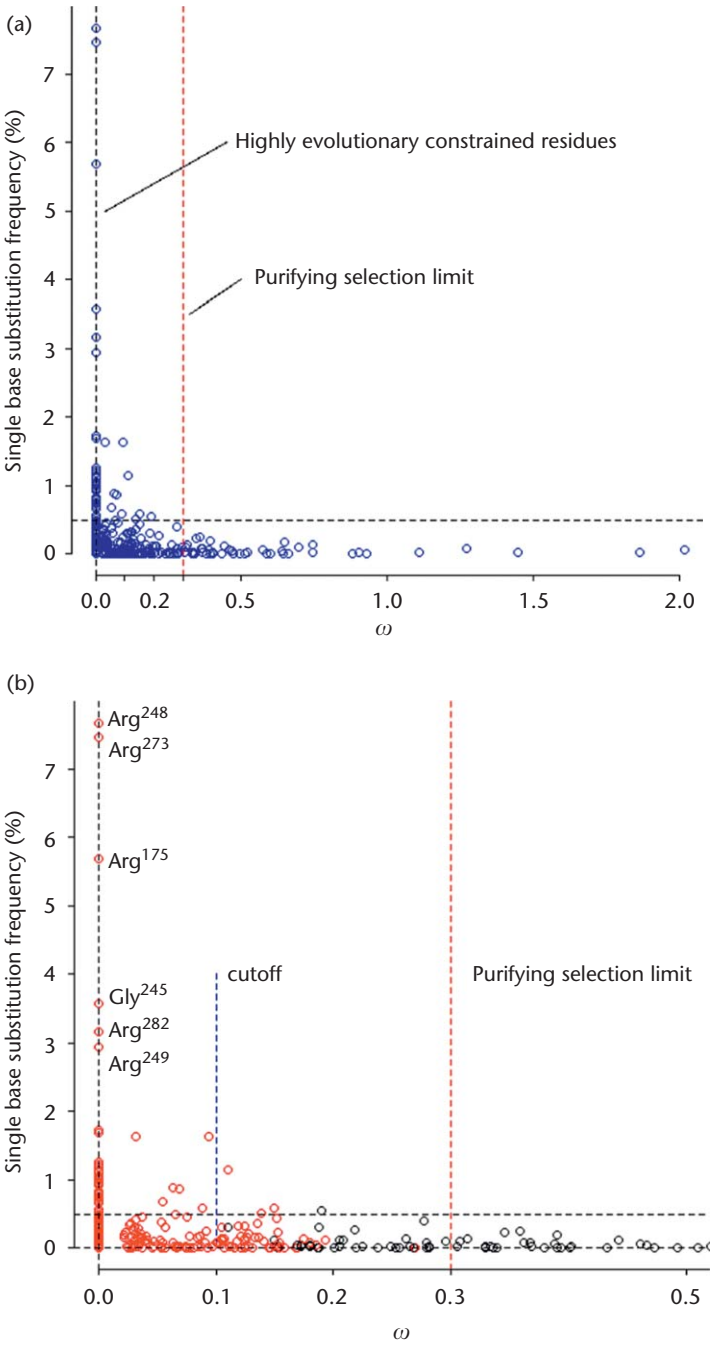

Figure 3 Mutation frequency and ω distribution in p53. (a) The distribution of p53 residues in the ω-frequency space describe an 'L-shaped' curve where sites under selective constraints (low ω values) are preferentially associated to high mutational frequencies associated to cancer. Conversely, residues above the limit of the effects of purifying selection (high ω) are preferentially associated to low mutational frequencies. (b) Mutational hotspots show high evolutionary constraints imposed by natural selection and the highest mutational frequencies. The cutoff value represents the maximum ω value for which residues were deduced to be under the influence of purifying selection at 95% or 99% confidence using the SLR method. This threshold represents the a priori hypothesis used to detect a statistical significance between mutation frequency and ω values using a large set of human disease genes. Reproduced from Arbiza *et al.* (2006), Copyright Elsevier (2006).

associated to the more common cancer mutations. **Table 4** shows the total number of mutations and the mean number of mutations per residue (numbers in bold) for the full protein, p53DB and p53TR domains, computed according to M8 and SLR models. In agreement with the earlier-mentioned expectation, the mean number of mutations per site (numbers in bold) shows a gradual increase according to the strength of natural selection according to the increase

Table 4 Cancer mutations and selective constraints

p53	Model	$\omega > 0.3$	$0.2 \leqslant \omega < 0.3$	$0.1 \leqslant \omega < 0.2$	$\omega < 0.1$	SLR[a]	ω_{M8}[b]	ω_{SLR}[c]	PC[d]
Full protein	M8	570	382	1714	15 165				
		9.5	**13.2**	**35.0**	**87.7**				
						16 883	13 028	12 992	13 152
						87.0	**119.5**	**120.3**	**120.6**
	SLR	430	250	1495	15 656				
		8.6	**11.4**	**25.3**	**87.0**				
	M8	437	337	1669	14 814				
DB		**23**	**25.9**	**50.6**	**113.1**				
						16 471	12 998	12 952	13 112
						99.2	**139.6**	**140.8**	**141.0**
	SLR	306	223	1436	15 292				
		20.4	**31.9**	**36.8**	**113.3**				
TD	M8	8	7	12	152				
		2.0	**2.3**	**2.4**	**7.6**				
						164	30	30	30
						6.3	**5**	**4.3**	**4.3**
	SLR	6	8	6	158				
		2.0	**2.5**	**1.5**	**7.5**				

Notes: Alternative classes of constraints collect a variable number of mutations associated to cancer (from the IARC TP53 R10 database) and a variable number of mutations per site (bold) in the protein. The increasing number of mutations per residue observed in ranges of ω with higher selective constraints ($0.3 > 0.2 > 0.1$) supports the hypothesis that natural selection works in proportion to the number of mutations in the population (see text). The class of phylogenetically conserved (PC) residues collects higher number of mutations per residue (120.6 and 141.0 for p53 and p53DB) highlighting its quality as a proxy for disease mutations. Similar values were observed for ω_{SLR} and ω_{SLR} classes. However, under the class of SLR a higher number of residues were observed associated to disease (194 and 166).

[a]Residues under the constraints of purifying selection evaluated by the SLR method at 95% and 99% statistical confidence.

[b]Residues with $\omega_{M8} \leqslant 0.033$.

[c]Residues with $\omega_{SLR} = 0$.

[d]Residues phylogenetically conserved throughout the p53 alignment.

on selective constraints from $\omega > 0.3$ to $\omega < 0.1$. The pattern is consistent throughout the whole protein, and is independent of the method used to estimate the ω values. The only exception occurred when considering the category where $0.1 \leqslant \omega < 0.2$ under the SLR method, but it seems justifiable given the low number of mutations observed in the p53TR domain. The category where residues are phylogenetically conserved (PC) showed the highest number of mutations producing disease per site in the full protein (120.6) which are very close to that estimated by $\omega_{SLR} = 0$ (120.3) and $\omega_{M8} < 0.033$ (119.5). This result points out the relevance of PC residues at the moment of defining amino acid sites where mutations producing disease are frequent. Alternatively, the SLR class of sites deducing purifying selection at 95% and 99% of confidence seems to be the more informative category when deducing the selective constraints imposed on the protein since it contains all the PC residues and shows that, in total, twice as many as those which are PC are selectively constrained with 95% and 99% of confidence. In addition, the SLR class seems to posses the ability to detect a greater number of mutations associated to disease per residue (87) for a greater proportion of residues.

Selective constraints, disease and polymorphism

In the previous section we predicted that mutations on sites carrying selective values lower than 0.1 ($\omega < 0.1$) are candidate sites to be associated to cancer in p53. Previously we demonstrated, using only 43 genes associated to disease, that this is the value that maximizes the differences ($p < < 0.001$) between mutations frequently associated and not associated to disease in humans (Arbiza *et al.*, 2006). Capriotti *et al.* (2008) suggested that disease and polymorphisms have alternative distribution of ω values. Using a large-scale testing set coming from the Swiss-Prot database (8987 amino acid variants, 6220 producing disease, 2767 polymorphic, in 1434 human proteins) they found a statistically significant association between high selective pressures and disease in contrast to low selective pressures and neutral polymorphic variants in human (**Figure 4**). These results suggest that disease-related protein variants and polymorphisms have significantly different evolutionary properties. The median ω value for disease-related protein variants was 0.072 lower than that for polymorphisms. This difference, although small, is very significant given a much larger distribution for ω values of

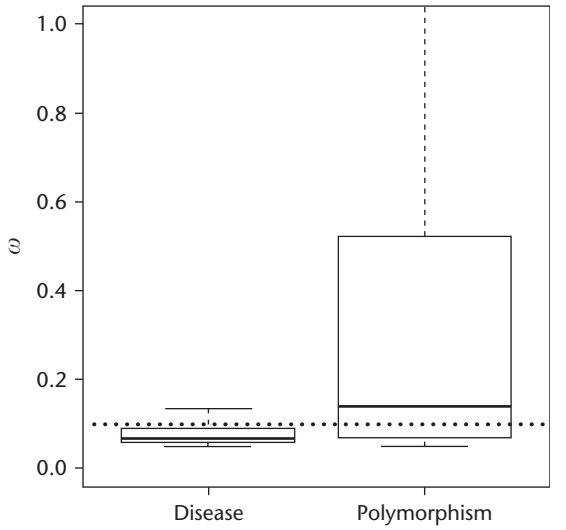

Figure 4 ω·distribution, disease and polymorphism. More than 8000 amino acid mutations defined as disease and polymorphic variation in the Swiss-Prot database are clearly differentiated by selective constraints. The boxplot shows the median (horizontal bold line), the upper and lower quartiles (box) and the interquartile range (dashed vertical lines). For visual clarity a horizontal dotted line indicates $\omega = 0.1$.

polymorphisms ($p = 2.2 \times 10^{-16}$). This result seems not only to confirm our previous prediction, but it also suggests that this evolutionary parameter is a feasible way to distinguish disease from polymorphism.

Bioinformatics Perspectives

Improving predictions using machine learning algorithms

Capriotti *et al*. (2008) have trained SVM classifiers using disease and polymorphic data from the Swiss-Prot database. Their results suggest that the application of SVMs classifiers outperforms previous bioinformatic results trying to infer disease mutations from sequence alignments and protein sequence information alone (like SIFT or PANTHER). **See also**: Bioinformatics; Neural Networks

They extend the implementation of sequence- and profile-based SVMs to include codon-based estimation of selective pressures at each position of the target sequences with noticeable success. The method proposed by Capriotti *et al*. (2008) (SeqProfCod, an acronym that comes from the use of information obtained from the human protein sequence, the profile of the alignment and codon selective pressures) achieves 82% overall accuracy, correctly predicting approximately 4% more protein variants than either SeqCod or SeqProf; two alternative SVMs specifically designed with sequence/codon and sequence/profile information only (**Table 5**). They demonstrate the synergy of combining two sources of information for predicting the functional effects of protein variants: protein sequence/profile-based information and the evolutionary estimation of the selective pressures at codon level. They finally suggest that the minimum error in the process of classification is reached when $\omega = 0.12$, which is very close to expectation in Arbiza *et al*. (2006). The results of large-scale application of SeqProfCod over all annotated point mutations in Swiss-Prot are available for download at http://sgu.bioinfo.cipf.es/services/Omidios/.

Table 5 Accuracy of the SVMs classifiers on the Swiss-Prot 2005 dataset

	Q(M)	Q(D)	Q(P)	*C*	AUC
SeqProf	0.78	0.80	0.74	0.52	0.85
SeqCod	0.79	0.82	0.74	0.53	0.86
SeqProfCod	0.82	0.84	0.77	0.59	0.88

Notes: Alternative SVMs classify disease and polymorphism with variable overall accuracy Q(M). However, when the information of selective constraints at codon level (Cod) are considered, the SVM increases the certainty for all the measured parameters. $Q(M) = 1/N(TP + TN)$, where TP and TN are true positive and negative predicted mutations and N the total number of mutations. $Q(s) = T(s)/(T(s) + F(s))$, where s is D: disease or P: polymorphism. C is the correlation coefficient and AUC the area under the ROC curve that represents the probability of correct classification over the whole range of cutoffs.

Computing constraints on the complete human genome

As was mentioned earlier, the population genetic analysis of human SNPs constitutes one of the most powerful tools to search for disease susceptibility genes (Collins *et al*., 2003). In this framework, the predicted functional effect of SNPs is gaining relevance as the selection criteria of alternative SNPs given that it constitutes a potentially relevant factor in significantly increasing the sensitivity of association tests (Botstein and Risch, 2003).

SNPs can be selected taking into account the evolutionary constraints of the region analysed along with its likelihood of being the causative agent of any type of damage. The PupaSuite web server (http://pupasuite.bioinfo.cipf.es) is a bioinformatic tool developed by Conde *et al*. (2006) for the selection of SNPs with potential phenotypic effect, specially oriented to help in the design of large-scale genotyping projects. This tool provides the ML estimation (M8 model and SLR method) of the evolutionary strength measured at a codon level for each one of the nsSNPs of the human genome (**Figure 5**). Users developing small- or large-scale genotyping analysis can select nsSNPs with high evolutionary constraints ($\omega \leqslant 0.12$) as possible candidates to successful experimental designs in the search for the genetic causes of human diseases. **See also**: Sequencing the Human Genome: Novel Insights into its Structure and Function

Conclusions

Selection against deleterious mutations (purifying selection) is accepted by most evolutionists as the predominant

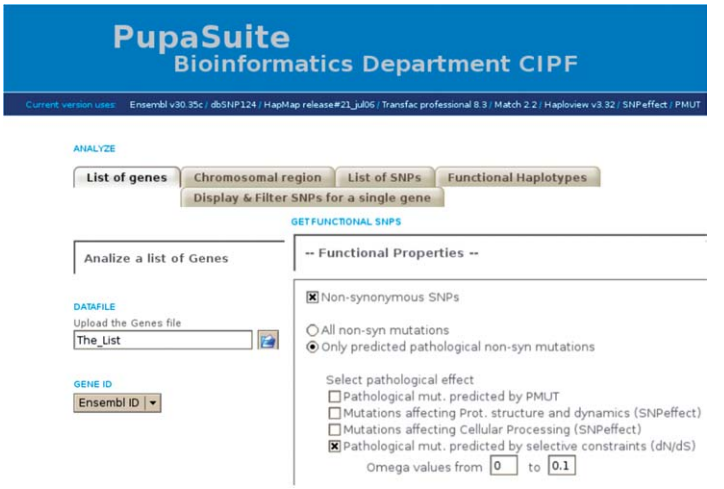

Figure 5 Analysis of the selective constraints of cSNPs in the human genome. The PupaSuite web server provides a complete set of tools for SNP characterization to assist users interested in genotyping experiments. The selection strength acting on all the cSNPs of the human genes can be reported according to different thresholds. By default the functional analysis of PupaSuite reports cSNPs showing $\omega < 0.1$.

form of selection at a molecular level. Earlier attempts at predicting functional consequences of nonsynonymous mutations represent indirect approaches to evaluate the strength of natural selection acting on polymorphic variation. Structural information alone, with or without the assistance of protein sequence alignments, was used in these methods as a multivariate proxy to deduce if a particular nsSNP produces a deleterious change in phenotype. The method discussed here differs from previous approaches through the explicit definition of the selective strengths occurring at a codon level. Codon-based ML models employed here make use of a number of parameters representing the more frequent and the more conserved changes occurring in the sequences during evolution. In this article we describe how an evolutionary parameter modelling biological sequences during millions of years allows to distinguish amino acid residues where human disease is frequent. This parameter allows differentiating significantly different distributions for disease and polymorphism. We hypothesize that nonsynonymous changes on amino acids showing $\omega < 0.1$ probably affect the normal function of proteins. The computation of this evolutionary parameter on all the coding sequences of the genome provides us with an a priori hypothesis of the phenotype effect of all the nsSNPs in the human genome.

Acknowledgement

This work is funded by Generalitat Valenciana GV06/080 and MEC BFU2006-15413-C02-02/BMC project. Many concepts, tables and figures of this manuscript were published in Arbiza *et al*. (2006), Copyright Elsevier (2006) and Capriotti *et al*. (2008), Copyright Wiley InterScience (2007).

References

Arbiza L, Dopazo J and Dopazo H (2006) Selective pressures at a codon-level predict deleterious mutations in human disease genes. *Journal of Molecular Biology* **358**: 1390–1404.

Bao L, Zhou M and Cui Y (2005) nsSNPAnalyzer: identifying disease-associated nonsynonymous single nucleotide polymorphisms. *Nucleic Acids Research* **33**: W480–W482.

Botstein D and Risch N (2003) Discovering genotypes underlying human phenotypes: past successes for Mendelian disease, future approaches for complex disease. *Nature Genetics* **33**(suppl.): 228–237.

Breiman L (2001) Random forest. Technical Report, Statistics Department UCB.

Bromberg Y and Rost B (2007) SNAP: predict effect of nonsynonymous polymorphisms on function. *Nucleic Acids Research* **35**(11): 3823–3835.

Capriotti E, Fariselli P, Calabrese R and Casadio R (2005) Predicting protein stability changes from sequences using support vector machines. *Bioinformatics* **21**(suppl. 2): ii54–ii58.

Capriotti E, Arbiza L, Casadio R *et al*. (2008) Use of estimated evolutionary strength at the codon level improves the prediction of disease-related protein mutations in human. *Human Mutation* **29**(1): 98–204.

Cargill M, Altshuler D, Ireland J *et al*. (1999) Characterization of single nucleotide polymorphisms in coding regions of human genes. *Nature Genetics* **22**: 231–238.

Chasman D and Adams RM (2001) Predicting the functional consequences of non-synonymous single nucleotide polymorphisms: structure-based assessment of amino acid variation. *Journal of Molecular Biology* **307**(2): 683–706.

Cho Y, Gorina S, Jeffrey PD and Pavletich NP (1994) Crystal structure of a p53 tumor suppressor-DNA complex: understanding tumorigenic mutations. *Science* **265**: 346–355.

Collins FS, Green ED, Guttmacher AE, Guyer MS and US National Human Genome Research Institute (2003) A vision for the future of genomics research. *Nature* **422**: 835–847.

Conde L, Vaquerizas JM, Dopazo H *et al*. (2006) PupaSuite: finding functional SNPs for large-scale genotyping purposes. *Nucleic Acids Research* **34**: W621–W625.

Felsenstein J (1985) Phylogenies and the comparative method. *The American Naturalist* **125**: 1–15.

Ferrer-Costa C, Gelpí JL, Zamakola L *et al*. (2005) PMUT: a web-based tool for the annotation of pathological mutations on proteins. *Bioinformatics* **21**(14): 3176–3178.

Golding B (1994) Using maximum likelihood to infer selection from phylogenies. In: Golding B (ed.) *Non-neutral Evolution. Theories and Molecular Data*, pp. 126–139. New York: Chapman & Hall.

IHMC (2007) A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**(7164): 851–861.

Karchin R, Diekhans M, Kelly L *et al*. (2005) LS-SNP: large-scale annotation of coding non-synonymous SNPs based on multiple information sources. *Bioinformatics* **21**(12): 2814–2820.

Kimura M (1983) *The Neutral Theory of Molecular Evolution*. New York: Cambridge University Press.

Massingham T and Goldman N (2005) Detecting amino acid sites under positive selection and purifying selection. *Genetics* **169**: 1753–1762.

Miller MP and Kumar S (2001) Understanding human disease mutations through the use of interspecific genetic variation. *Human Molecular Genetics* **10**: 2319–2328.

Ng PC and Henikoff S (2001) Predicting deleterious amino acid substitutions. *Genome Research* **11**(5): 863–874.

Ng PC and Henikoff S (2003) SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Research* **31**: 3812–3814.

Ramensky V, Bork P and Sunyaev S (2002) Human non-synonymous SNPs: server and survey. *Nucleic Acids Research* **30**(17): 3894–3900.

Reumers J, Maurer-Stroh S, Schymkowitz J and Rousseau F (2006) SNPeffect v2.0: a new step in investigating the molecular phenotypic effects of human non-synonymous SNPs. *Bioinformatics* **22**: 2183–2185.

Santibáñez Koref MF, Gangeswaran R, Santibáñez Koref IP, Shanahan N and Hancock JM (2003) A phylogenetic approach to assessing the significance of missense mutations in disease genes. *Human Mutation* **22**: 51–58.

Saunders CT and Baker D (2002) Evaluation of structural and evolutionary contributions to deleterious mutation prediction. *Journal of Molecular Biology* **322**(4): 891–901.

Sunyaev S, Ramensky V, Koch I *et al*. (2001) Prediction of deleterious human alleles. *Human Molecular Genetics* **10**: 591–597.

Sunyaev SR, Lathe WC 3rd, Ramensky VE and Bork P (2000) SNP frequencies in human genes: an excess of rare alleles and differing modes of selection. *Trends in Genetics* **16**(8): 335–337.

Thomas PD and Kejariwalet A (2004) Coding single-nucleotide polymorphisms associated with complex vs. Mendelian disease: evolutionary evidence for differences in molecular effects. *Proceedings of the National Academy of Sciences of the USA* **101**(43): 15398–15403.

Thomas PD, Campbell MJ, Kejariwal A *et al*. (2003) PANTHER: a library of protein families and subfamilies indexed by function. *Genome Research* **13**: 2129–2141.

Wang Z and Moult J (2001) SNPs, protein structure, and disease. *Human Mutation* **17**: 236–270.

Yang Z (2003) Adaptive molecular evolution. In: Balding D, Bishop M and Cannings C (eds) *Handbook of Statistical Genetics*, 2nd edn, pp. 229–254. New York: Wiley.

Yang Z (2007) PAML 4: phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution* **24**(8): 1586–1591.

Ye ZQ, Zhao SQ, Gao G *et al*. (2007) Finding new structural and sequence attributes to predict possible disease association of single amino acid polymorphism (SAP). *Bioinformatics* **23**(12): 1444–1450.

Yue P, Melamud E and Moult J (2006) SNPs3D: candidate gene and SNP selection for association studies. *BMC Bioinformatics* **7**(1): 166.

Zuckerkandl E and Pauling L (1965) Molecules as documents of evolutionary history. *Journal of Theoretical Biology* **8**: 357–366.

Further Reading

Ng and Henikoff (2006) Predicting the effects of amino acid substitutions on protein function. *Annual Review of Genomics and Human Genetics* **7**: 61–80.

Yampolsky LY, Kondrashov FA and Kondrashov AS (2005) Distribution of the strength of selection against amino acid replacements in human proteins. *Human Molecular Genetics* **14**: 3191–3201.