METHOD

Human Mutation

OFFICIAL JOURNAL
HGVS
HUMAN GENOME VARIATION SOCIETY
www.hgvs.org

Functional Annotations Improve the Predictive Score of Human Disease-Related Mutations in Proteins

Remo Calabrese, Emidio Capriotti, Piero Fariselli, Pier Luigi Martelli, and Rita Casadio*

Laboratory of Biocomputing, CIRB/Department of Biology, University of Bologna, Bologna, Italy

Communicated by Mauno Vihinen

Received 16 December 2008; accepted revised manuscript 1 May 2009.
Published online 12 May 2009 in Wiley InterScience (www.interscience.wiley.com). DOI 10.1002/humu.21047

ABSTRACT: Single nucleotide polymorphisms (SNPs) are the simplest and most frequent form of human DNA variation, also valuable as genetic markers of disease susceptibility. The most investigated SNPs are missense mutations resulting in residue substitutions in the protein. Here we propose SNPs&GO, an accurate method that, starting from a protein sequence, can predict whether a mutation is disease related or not by exploiting the protein functional annotation. The scoring efficiency of SNPs&GO is as high as 82%, with a Matthews correlation coefficient equal to 0.63 over a wide set of annotated nonsynonymous mutations in proteins, including 16,330 disease-related and 17,432 neutral polymorphisms. SNPs&GO collects in unique framework information derived from protein sequence, evolutionary information, and function as encoded in the Gene Ontology terms, and outperforms other available predictive methods.
Hum Mutat 30, 1237–1244, 2009. © 2009 Wiley-Liss, Inc.

KEY WORDS: missense mutation; support vector machine; Gene Ontology; disease-related SNP

Introduction

Recent estimates indicates that single nucleotide polymorphisms (SNPs) occur approximately every 200 bases in DNA of human populations (Ensembl release 53.36o). SNPs have been correlated to human evolution, drug sensitivity, and disease susceptibility [Barbujani and Goldstein, 2004; Bell, 2004; Edmonds et al., 2004; Goldstein and Cavalleri, 2005; Ng and Henikoff, 2002; Robert et al., 2005]. The international and ongoing HapMap (http://www.hapmap.org) project is funded to determine the common patterns of DNA sequence variation in the human genome and its relation to common diseases [Cotton et al., 2008; Wang et al., 1998]. In recent years, new experimental techniques for large-scale SNP identification in the human population have allowed the exponential increase of the dbSNP database (http://www.ncbi.nlm.nih.gov/projects/SNP/) that presently contains over 10 million validated cases (dbSNP 129) [Sherry et al., 2001]. Other important and human devoted databases include the Human Gene Mutation Database (HGMD, http://www.hgmd.cf.ac.uk/ac/index.php) and the Human Genome Variation database of Genotype-Phenotype (HGVbase2GP, http://www.hgvbaseg2p.org/index).

A major problem is SNP annotation. To this end several Web servers and tools have been developed to relate SNPs to potential phenotypic effects. Although OMIM (http://www.ncbi.nlm.nih.-gov/omim/) is the major source of data, annotation is obtained with different approaches. SNPs can be retrieved and analyzed [Riva and Kohane, 2002; Stitziel et al., 2004; Wang et al., 2005] also with links to the Gene Ontology (GO) sequence-associated terms [Schwarz et al., 2008].

Different computational tools were developed to study the effects of SNPs in the human genome at the phenotype level. Correlations among DNA mutations stored in the databases and the insurgence of pathologic phenomena were computed [Collins et al., 1997; Rish and Merikangas, 1996]. In general, mutations occurring in coding regions may have a greater impact on the gene functionality than those occurring in noncoding regions [Cargill et al., 1999].

In this article we focus on missense nonsynonymous SNPs, that is, those that change single residues in the protein sequence. Several computational methods are available to predict when a mutation is disease related, starting from the protein sequence and/or protein multiple sequence alignments (for a recent review, see [Tavtigian et al., 2008]). They are based on: (1) sequence homology [Ng and Henikoff, 2002, 2003; Thomas et al., 2003], (2) empirical rules [Ramensky et al., 2002; Sunyaev et al., 2001], (3) structural criteria [Chasman and Adams, 2001; Cheng et al., 2008; Wang and Moult, 2001; Worth et al., 2007; Yue and Moult, 2005, 2006], (4) artificial neural networks [Bromberg and Rost, 2007; Ferrer-Costa et al., 2002, 2004, 2005], (5) decision trees [Dobson et al., 2006; Krishnan and Westhead, 2003], (6) random forests [Bao and Cui, 2005], and (7) support vector machines (SVMs) [Capriotti et al., 2006; Kulkarni et al., 2008; Tian J et al., 2007]. All these approaches exploit information from the protein sequence, the sequence profile, the three dimensional structure (3D) of the protein and/or adopt some combinations of them. The structural approach is affected by limited availability of 3D data so that it is not generally applicable. Evolutionary information as encoded in the sequence profile is the most important piece of information for improving the predictive performance, as indicated by the results of PANTHER (Protein ANalysis THrough Evolutionary Relationships) [Thomas and Kejariwal, 2004], SIFT (Sorting Intolerant From Tolerant) [Ng and Henikoff, 2003], PolyPhen (Polymorphism Phenotyping) [Ramensky et al., 2002], and other predictors described in the literature [Bromberg and Rost, 2007], including ours [Capriotti et al., 2006]. When the computed selective pressure of the mutation at the codon level [Arbiza et al., 2006] was cast into the method, some predictive improvement was detected [Capriotti et al., 2008].

Emidio Capriotti is now at the Genomics Unit, Department of Bioinformatics, Centro de Investigacion Principe Felipe (CIPF), Autopista del Saler 16, 46013 Valencia, Spain.

*Correspondence to: Rita Casadio, Laboratory of Biocomputing, CIRB/Department of Biology, University of Bologna, via Irnerio 42, 40126 Bologna, Italy. E-mail: casadio@biocomp.unibo.it

An alternative source of information that can provide useful indications is the GO database that addresses the need of coherent descriptions of gene products [Ashburner et al., 2000]. The GO project has developed tree-structured and controlled vocabularies (ontology) that describe gene products in terms of their associated biological processes, cellular components, and molecular functions. We therefore compute a GO-based score to increase the information provided to our new implementation. The adoption of a similar GO-based score was previously introduced by other authors to predict single point protein mutations focusing on cancer disease in combination with other features (PFAM- and SIFT-derived scores) [Kaminker et al., 2007a,b].

In this article we describe a tool for predicting whether human SNPs are or are not disease-associated by including the protein sequence GO terms in our method (SNPs&GO). Our method is a robust predictor, trained/tested over more than 33,000 mutations that casts in a unique input vector various features, including sequence information, evolutionary information derived in different ways, and our defined functional GO score. The efficiency of the prediction is quite high as indicated by different scoring indexes and by the comparison with other previously implemented predictors. The overall SNPs&GO accuracy and Matthews correlation coefficient values, computed by adopting a cross validation procedure on a recent human SNPs data set [Boeckmann et al., 2003], are as high as 0.82 and 0.63, respectively.

Materials and Methods

The Mutation Data Set

Our SNPs data set is derived from release 55.2 (April 2008) of the Swiss-Prot database [Boeckmann et al., 2003]. The choice of the training data set is particularly critical when developing machine learning methods. The issue was recently addressed [Care et al., 2007], and it was found that largely irrelevant rules may be derived for missense SNPs (mSNPs) predictions from mutagenesis data [Chasman and Adams, 2001; Krishnan and Westhead, 2003] and from the generation of neutral data starting from pseudomutations between orthologous proteins [Ferrer-Costa et al., 2002, 2004, 2005; Sunyaev et al., 2001; Yue and Moult, 2005, 2006]. The best data set for human mSNP predictions is that of the Swiss-Prot annotated variants [Care et al., 2007; Yip et al., 2004]. We retrieved our data set from Swiss-Prot with the following constraints: (1) the protein source is *Homo sapiens*, (2) the mutations are related to diseases or neutral polymorphisms (no unclassified cases are considered), (3) the data are relative to single-point protein mutations (no deletion or insertion mutations are taken into account).

After this selection procedure, we filtered out neutral polymorphisms belonging to proteins of class I and II of the Major Histocompatibility complex (human leukocyte antigen, HLA), because these proteins are naturally hypervariable. We ended up with a data set (SP_human) consisting of 33,762 different single point mutations (16,330 of which are disease-related and 17,432 are described as neutral polymorphisms), obtained from 7,265 protein sequences.

Implementation of the Predictors

Our task is to predict whether a given single point protein mutation is a neutral polymorphism or if it is involved with the insurgence of a human genetic disease. The task can be cast as a classification problem for the protein upon mutation. The SVM classifies mutations into disease-related (desired output set to 0) and neutral polymorphism (desired output set to 1). No attempts were made to improve the accuracy, changing the decision threshold that is set equal to 0.5. To develop a more accurate tool we include: the local sequence environment of the mutation at hand, features derived from sequence alignment, prediction data provided by the PANTHER classification system and a functional-based log-odds score calculated considering the GO classification (Fig. 1). The final input vector consists of 52 values:

- 40 components encode for the mutations and sequence local information (Seq);
- four inputs concern features derived from sequence profile plus an extra one (a bit) codifying the presence/absence of the features themselves (Prof);
- four values represent selected parameters of PANTHER (prediction output plus an extra node encoding the presence/absence of PANTHER output) (PANTHER);

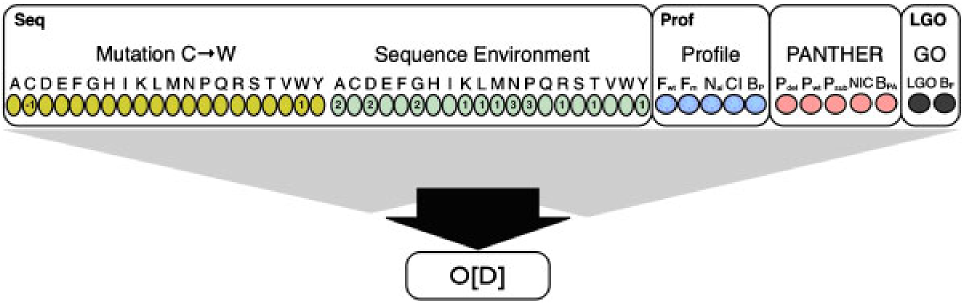
Seq
Mutation C→W
Sequence Environment
A C D E F G H I K L M N P Q R S T V W Y
-1
1
A C D E F G H I K L M N P Q R S T V W Y
2
2
2
1
1
1
3
3
1
1
1
Prof
Profile
F_{wt} F_m N_{al} CI B_P
PANTHER
P_{del} P_{wt} P_{sub} NIC B_{PA}
LGO
GO
LGO B_f
O[D]

Figure 1. SNPs&GO input schema. Different features from different metrics are encoded and cast in an unique framework. A first level of input takes into consideration sequence-derived information (Seq): the first 20 values encode the involved mutation (e.g., C->W) setting the wild-type residue to -1 and the mutated residue to 1(yellow nodes); the following 20 values (grey nodes) describe the mutation local environment in terms of occurrence of residues (nodes containing 1) inside a window centred on the mutation at hand. A second level of information encodes the sequence profile (Prof, five nodes in blue): F_{wt} is the frequency of wild-type residue, F_m is the frequency of the mutated residue, N_{al} is the number of aligned sequences at the mutated position, CI is a conservation index, B_p is a bit value related to the presence/absence of the sequence profile. The information derived from PANTHER output is also encoded (PANTHER, five nodes in pink): P_{del} is the disease-related probability of the mutation at hand, P_{wt} is the probability of the wild-type residue, P_{sub} is the probability of the mutated residue, NIC is Number of Independent Counts, B_{PA} is also a presence/absence bit. The last two nodes (black) encode GO terms: LGO is the log-odds score derived from the GO database and its occurrence (B_f). For more details see Materials and Methods and the text (Table 1). Each level of information is given as input to a support vector machine (SVM) separately or combined, for a total of 10 different SVM predictors. The input vector of SNPs&GO includes all the information is the input vector of SNPs&GO.

- two components encode for the GO log-odd score (LGO) and for its presence/absence (LGO).

Each of the four sets of input values coding for sequence, profile, PANTHER-derived, and GO-derived information, were first implemented independently (Seq, Prof, PANTHER, LGO in Table 1); then integrated step by step as indicated when necessary.

For the SVM implementation we use the LIBSVM library (http://www.csie.ntu.edu.tw/~cjlin/libsvm/) (with a RBF kernel function, $K(x_i,x_j) = \exp(-G\|x_i - x_j\|^2)$) [Chang and Lin, 2001]. To increase the generalization performance of the method we optimized the two critic parameters: C and γ. The C parameter concerns the penalty with which the classifier is allowed to make errors in training/testing phases (soft margin), whereas the γ parameter is an intrinsic value of RBF kernel concerning the width of support vector. LIBSVM offers the possibility to perform an automatic grid search for the above parameters. The optimal values of the parameters are $C = 8$ and $\gamma = 0.03125$.

Scoring the Performance

The results obtained with our SVM methods are evaluated using a crossvalidation procedure on the SP_human data set. The reported data for the classification task performed by the SVM methods are obtained adopting a 20-fold crossvalidation procedure in such a way that the ratio of the disease-related to the neutral polymorphism mutations corresponds to the original distribution of the whole set. Proteins in different crossvalidation sets share less than 30% sequence identity. Furthermore, all the proteins in the SP_human data sets are clustered according to their sequence similarity with the blastclust program in the BLAST suite (by adopting the default value of length coverage equal to 0.9 and the percentage similarity threshold equal to 30%) [Altschul et al., 1997]. We kept the mutations detected on the same cluster of protein sequences in the same training set to prevent an overestimation of the results. Performance is scored with several measures. For sake of completeness here we review the ones adopted in this article. The efficiency of the predictor is scored using the statistical indexes defined in the following.

Table 1. Different SVM Implementations Improve the Predictive Performance

Method	Q2	P[D]	Q[D]	P[N]	Q[N]	C	AUC
Seq	0.68	0.68	0.64	0.69	0.72	0.36	0.75
Prof	0.69	0.68	0.76	0.79	0.84	0.38	0.77
PANTHER	0.74	0.77	0.73	0.71	0.76	0.48	0.82
LGO	0.68	0.74	0.52	0.65	0.83	0.37	0.79
Seq+Prof	0.76	0.75	0.73	0.76	0.78	0.51	0.82
Seq+PANTHER	0.77	0.77	0.73	0.77	0.80	0.53	0.84
Seq+LGO	0.75	0.78	0.68	0.73	0.82	0.51	0.84
Seq+Prof+PANTHER	0.78	0.78	0.75	0.78	0.81	0.56	0.85
Seq+Prof+LGO	0.80	0.81	0.76	0.79	0.84	0.60	0.88
Seq+PANTHER+LGO	0.80	0.82	0.76	0.79	0.84	0.60	0.88
SNPs&GO	0.82	0.83	0.78	0.80	0.85	0.63	0.89

Scoring indexes are evaluated for single point protein mutations related to human disease (D) and neutral polymorphism (N), respectively; Q2 is the overall accuracy; Q is the coverage of the class (D, N); P is the probability of correct predictions (or accuracy) of the class (D, N); C is the Matthews Correlation Coefficient; AUC is the area under the ROC Curve (for a mathematical definition of the different indexes see the Materials and Methods section). Different SVMs are analyzed: Seq is based only on mutation and sequence environment information; LGO is based only on Gene Ontology derived log-odds score; Prof exploits only information derived from sequence profile; PANTHER exploits only features derived from PANTHER output; Seq+Prof is the SVM based on mutation, sequence and sequence profile information; Seq+PANTHER is the SVM based on mutation, sequence environment and PANTHER output features; Seq+LGO is the SVM based on mutation, sequence environment and Gene Ontolgy derived log-odds score; Seq+Prof+ PANTHER, Seq+Prof+LGO, Seq+PANTHER+LGO, are the SVMs combining different sources of information coming from the above metrics. The complete SNP&GO predictor combines all the available sources of information: Seq, Prof, PANTHER, LGO. The absolute standard deviations of the listed average values of the different scoring indexes (after the cross validation procedure) ranged from 0.02 to 0.07.

The overall accuracy is:

$$Q2 = P/N \tag{1}$$

where P is the total number of correctly predicted mutations and N is the total number of mutations.

The correlation coefficient C is defined as:

$$C(s) = [p(s)n(s) - u(s)o(s)]/D \tag{2}$$

where D is the normalization factor:

$$D = [(p(s) + u(s))(p(s) + o(s))(n(s) + u(s))(n(s) + o(s))]^{1/2} \tag{3}$$

for each class s (D and N, for disease-related and neutral polymorphism, respectively); $p(s)$ and $n(s)$ are the total number of correct predictions and correctly rejected assignments, respectively, and $u(s)$ and $o(s)$ are the numbers of under- and over-predictions.

The coverage (sensitivity) for each discriminated class s is evaluated as:

$$Q(s) = p(s)/[p(s) + u(s)] \tag{4}$$

where $p(s)$ and $u(s)$ are the same as in Equation 3.

The probability of correct predictions $P(s)$ (or accuracy for s, or specificity) is computed as:

$$P(s) = p(s)/[p(s) + o(s)] \tag{5}$$

where $p(s)$ and $o(s)$ are the same as in Equation 3 (ranging from 0 to 1).

Finally, it is very important to assign a reliability score to each prediction. For each output O(D) computed by the SVM for the category Disease and indicating the probability of being disease-associated, the reliability score (RI) is obtained by computing:

$$\mathrm{RI} = 20 * |\mathrm{O(D)} - 0.5| \tag{6}$$

Other standard scoring measures, such as the area under the ROC curve (AUC) and the true positive rate [TPR = $Q(s)$] at 5% of false positive rate [FPR = $1-P(s)$] are also computed [Baldi et al., 2001].

Encoding Sequence Information

The input vector portion relative to sequence information consists of 40 values: the first 20 (the 20 residue types) explicitly define the mutation by setting to -1 the element corresponding to the wild-type residue, and to 1 the newly introduced residue (all the remaining elements are kept equal to 0). The last 20 input values encode for the mutation sequence environment (again, the 20 elements represent the 20 residue types). Each input is provided as the number of the encoded residue type, found inside a window centred at the residue that undergoes mutation and that symmetrically spans the sequence to the left (N-terminus) and to the right (C-terminus) with a length of 19 residues (Fig. 1; for more details see [Capriotti et al., 2006]). The sequence environment inputs sum to 18, not including the central residue.

Encoding Profile Information

We derive for each mutation: the frequency of the wild-type (F_{wt}), the frequency of the mutated residue (F_m), the number of aligned sequences (N_{al}), and a conservation index (CI) for the position at hand: the more a residue is functionally important the more it is conserved during evolution [Pei and Grishin, 2001]

(Fig. 1). The conservation index is calculated as:

$$\mathrm{CI}(i) = \left[\sum_{a=1}^{20}(f_a(i) - f_a)^2\right]^{1/2} \tag{7}$$

where $f_a(i)$ is the relative frequency of residue a at mutated position i and f_a is the overall frequency of the same residue in the alignment. The sequence profile is computed from the output of the BLAST program, running on the uniref90 database (release 13.3 April 2008) (E-value threshold = $1e^{-9}$, number of runs = 1). When sequence profile values at position i are missing we set the frequencies of wild-type and mutant residue equal to 0.5, the number of aligned sequence equal to 2; CI is calculated accordingly and the extra bit value is set to 0.

Encoding the PANTHER Outputs

PANTHER consists of two main components: the PANTHER library and the PANTHER index. The first one is a collection of "books" each representing a protein family as a multiple sequence alignment, a Hidden Markov Model (HMM) and a family tree. The second one is a slim ontology for describing molecular functions and biological processes associated with the families. There is a PANTHER application (csnpAnalysis1.0) that uses the HMM family to classify mSNPs, according to their likelihood of affecting the protein function. Because evolutionary-related sequences are used to estimate the probability of a given residue at a particular position in a protein, the method can be referred to as generating "position-specific evolutionary conservation" (PSEC) scores. For our purposes, we extracted four features from the PANTHER outputs to implement the "meta-predictor": the probability of a mSNP being disease related (P_{del}), the probability of the wild-type residue (P_{wt}), the probability of substituted/mutant amino acid (P_{sub}), and the number of independent counts (NIC), which is a measure of the global diversity of the sequences over which a position has been conserved. When PANTHER output features at position k are missing we adopted an a priori guess of neutrality for mutations not scored, setting the values of interest equal to: $P_{del} = 0.5$, $P_{wt} = 0.05$, $P_{sub} = 0.05$, and NIC = 1 (bit = 0).

Computing the LGO Score

The GO log-odds score (LGO) is computed to derive information related to the correlation among a given mutation type (disease-related and neutral) and the protein function. The annotation data are relative to the GO Database version 1.37 and are retrieved at the Web resource hosted at European Bionformatics Institute (EBI): www.geneontology.org/cgi-bin/downloadGOGA.pl/gene_association.goa_human.gz. To compute LGO, first we derived the GO terms (from all the three branches: Molecular Function, Biological Process and Cellular components, when available) for all our proteins in the data set (SP_human). For each annotated term the appropriate ontology tree was traversed upward to retrieve all the parent terms with the GO-TermFinder-0.8 tool (http://search.cpan.org/dist/GO-TermFinder/) [Boyle et al., 2004] and taking care of introducing a GO term only once. The information is then derived by computing a log-odds score associated to each protein:

$$\mathrm{LGO} = \Sigma \log_2[(\mathrm{GTO(GO)_D} + 1)/(\mathrm{GTO(GO)_N} + 1)] \tag{8}$$

where GTO (Gene ontology Term frequency of Occurrence) is the frequency of occurrence of a given GO term for the disease-related (D) and neutral mutations (N); the pseudocount 1 is added to avoid undetermined ratios.

The LGOs are evaluated considering GTO values computed over the training sets without including in the counts the GO terms of the corresponding test set. This is done to crossvalidate also this type of information.

Results

The Performance of SNPs&GO

To relate a mutation in a protein sequence to a disease, we may take advantage of what machine learning and statistical methods have taught us in the past 10 years or so of sequence analysis. Among different computational methods, classifiers based on support vector machines are among the most powerful for their classification capabilities [Bishop, 2006]. Also, evolutionary information, as derived from a sequence profile of the target sequence to its homologs in the sequence databases, is of fundamental importance for detecting mutations that affect human health [Ramensky et al., 2002].

The main novelty described in this article with respect to previous applications is the use of functional GO terms (Fig. 1). An alternative way to include evolutionary information is to exploit HMMs computed on specific protein families. For this reason we adopt some features from the output of the PANTHER Classification System, a unique resource that classifies genes by their functions with HMMs [Thomas and Kejariwal, 2004].

The new input takes into account in a unique vector (Fig. 1) different features derived from: (1) the mutation type and from the local sequence environment where the mutation occurs (Seq). The relevance of the inclusion of sequence information and sequence profile was previously discussed [Capriotti et al., 2006]; (2) the sequence profile (Prof) [by evaluating the frequency of the wild-type residue (F_{wt}), of the mutated residue (F_m), the number of sequence in the alignment (N_{al}), and a conservation index (CI, Equation 7)]; (3) the PANTHER outputs (PANTHER) (the P_{del}; the occurrence of P_{wt}; the P_{sub}; the PANTHER measure of the global diversity of the sequences over which a position is conserved (NIC)); (4) the GO functional annotation system (after computing a LGO; Equation 8).

Our new method is called SNPs&GO (classifying human SNPs by including GO), and its necessity is demonstrated by implementing different SVM predictors with an increasing level of input complexity (Table 1). The training/testing set of mutations (SP_human) was derived from the Swiss-Prot database (release April 2008) (detailed in Materials and Methods) and comprises 33,762 mutations from 7,265 proteins, including neutral and disease-related mutations (17,432 and 16,330, respectively). In Table 1 we list the results that are obtained adopting a crossvalidation procedure under different implementation conditions on the same training/testing set. Performance is measured by computing different scoring indexes: Q2, the overall accuracy; P(D), the rate of correct predictions for the disease-related mutations (D); Q(D), the coverage (number of correctly predicted mutations) for the disease-related mutations; P(N), the rate of correct predictions for the neutral mutations (N); Q(N), the coverage for the neutral mutations; AUC, an estimate of how the predictor is different from a random predictor characterised by AUC = 0.5 (for a detailed definitions of the indexes see Materials and Methods).

From this effort it can be concluded that the more sources of information are collected the higher the performance scores are. For a more general evaluation of the performance of SNPs&GO, we also computed the receiver operating characteristic (ROC) curves of the different SVMs and calculated the relative AUCs (see the rightmost column of Table 1; AUC, and Fig. 2). From all the index values, it is evident that SNPs&GO is endowed with a better predictive performance than the other SVMs with less input information, scoring with values of Matthews correlation coefficient (C) and accuracy (Q2) 27 and 14 percentage points higher than Seq, the basic predictor that implements only sequence information. Also, the SNPs&GO area under the ROC curve value (AUC) is 14 percentage points higher than that of Seq (Fig. 2).

From Table 1 it is evident that the implementation of LGO on top of the basic predictors, as well as on top of their respective combination, promotes a better performance.

Specifically, the implementation of LGO on top of Seq+Prof+PANTHER improves Q2 and C values by 0.04 and 0.07, respectively, indicating that the addition of functional information is crucial for improving the scoring indexes. For a given mutation in the protein sequence SNPs&GO returns the probability value of being disease-related or not. The value is then used to calculate the Reliability Index (RI) (see Materials and Methods). In Figure 3, Q2 and C values are plotted as a function of the RI values. From this, it appears that when the RI value is 5, over 70% of the SP_human database can still be predicted with Q2 and C values equal to 0.9 and 0.74, respectively.

Recently we described a SVM implementation that includes the evolutionary selective pressure as an additional piece of information to the input code (SeqProfCod [Capriotti et al., 2008]). The evaluation of the evolutionary selective pressure is done at the codon level and it needs to be precomputed given its complexity and computational cost. To assess the efficacy of the GO terms we therefore tested SeqProfCod and SNPs&GO on a subset of mutations for which the evolutionary selective pressure values were made available after computing with the procedure previously described (Capriotti et al., 2008, http://sgu.bioinfo.cipf.es/services/Omidios/). The testing subset was some 28% of the all data set of mutations described in this article, containing 9,544 mutations (in 1,826 proteins), 6,282 of which are disease-related. It should be also stressed that all the different implementations during the experiment were used adopting a rigorous crossvalidation procedure. The data are shown in Table 2. It is evident that the addition of different features only at the protein level, including GO terms (SNPs&GO), are sufficient to obtain a score higher than SeqProfCod (first three rows in Table 2). Furthermore, the addition of the information derived from the evolutionary selective pressure computed at the codon level (+Cod) on top of SNPs&GO (SNPs&GO+Cod) is not significantly increasing the predictive performance. For this reason, we focus on SNPs&GO that includes the information that can be computed at the sequence level. For sake of comparison we also tested our predictor on previous sets adopted to score SeqProfCod. The results (last four rows in Table 2) confirm the improvement obtained by exploiting the information of the GO terms.

A Benchmark with Other Methods

We compare our method with others, commonly used, available in the Web and downloadable for in-house implementation such as SIFT [Ng and Henikoff, 2003] and PANTHER [Thomas et al., 2003], or making available precalculated data such as PolyPhen [Ramensky et al., 2002] and Eremorph [Kulkarni et al., 2008] (Table 3). PolyPhen is based on a decision tree, and takes into account information derived by structural parameters, functional annotations, and sequence alignments. SIFT makes predictions on the basis of sequence homology, while PANTHER exploits HMM models of protein families and evaluates the residue conservation in a given protein family. Eremorph exploits different parameters of interspecies nucleotide conservation coupled with the BLOSUM [Henikoff and Henikoff, 1992] and PAM [Dayhoff et al., 1978] substitution matrix-scoring indexes. As indicated in the rightmost column of Table 3, not all the predictors are able to make predictions for the entire data set. We also compare SNPs&GO with our previous method (HybridMeth [Capriotti et al., 2006]) based only on sequence and profile information. It appears that SNPs&GO outperforms all previous implementations, including ours. The benchmarking also indicates that one of the advantages of our present implementation is the possibility of predicting all the mutations of the data set in real time and in an interactive way (see the Web site of SNPs&GO at http://snps-and-go.biocomp.unibo.it).

Scoring the Prediction in Relation to Different Morbidities

In Table 4 the ability of the predictor is scored (Q[D], coverage) with respect to the number of mutations in database related to a

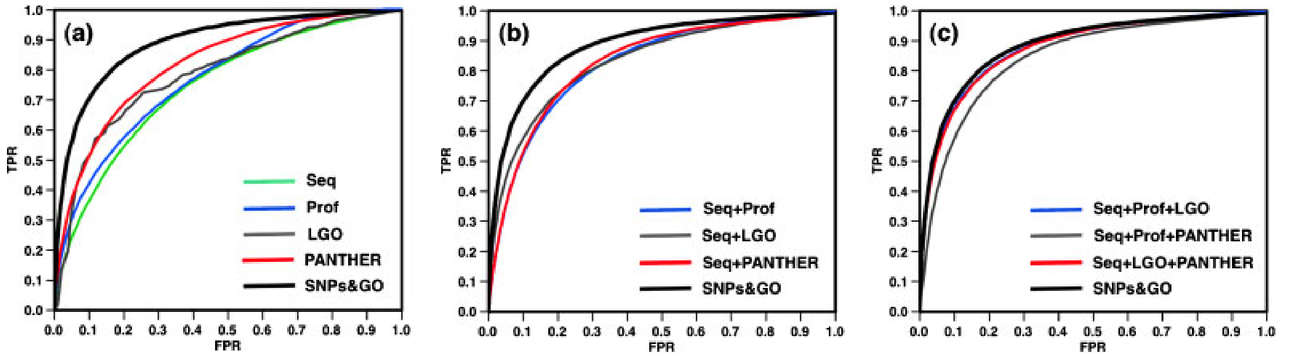

Figure 2. The ROC curves of the different predictors. The ROC curve of the SNPs&GO method is compared with those of the other predictors whose performance is reported in Table 1. (**a**) SNPs&GO versus Seq, Prof, LGO, and PANTHER. (**b**) SNPs&GO versus Seq+Prof, Seq+LGO, Seq+PANTHER. (**c**) SNPs&GO versus Seq+Prof+LGO, Seq+Prof+PANTHER, Seq+LGO+PANTHER. FPR and TPR are the false and the true positive rates, respectively (see Materials and Methods for definition).

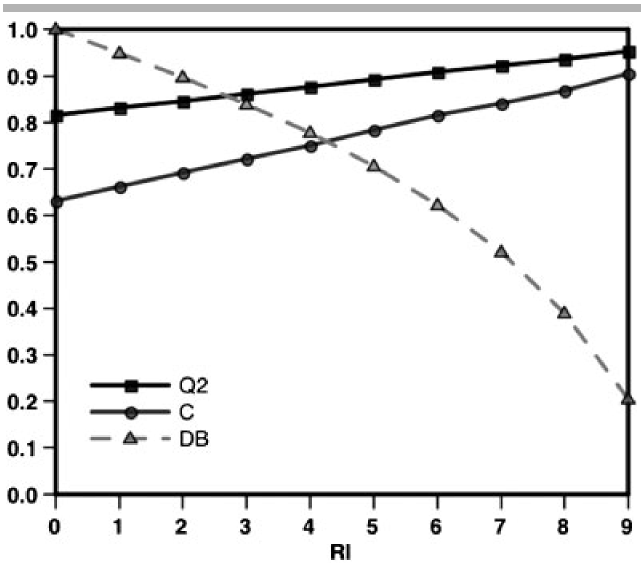

Figure 3. Overall accuracy and Matthews correlation coefficient of SNPs&GO as a function of Reliability Index (RI). Overall accuracy (Q2) and Matthews correlation coefficient (C) of SNPs & GO are plotted as a function of the RI of the prediction. DB is the fraction of SP_human data set with RI values ≥ of a given threshold.

Table 2. Go Terms versus Evolutionary Selective Pressure at the Codon Level

Method	Q2	P[D]	Q[D]	P[N]	Q[N]	C	Dataset
SeqProfCod[a]	0.76	0.83	0.80	0.64	0.69	0.48	Intersection
SNPs&GO[b]	0.82	0.84	0.89	0.76	0.68	0.59	Intersection
SNPs&GO+Cod[c]	0.83	0.85	0.89	0.77	0.70	0.60	Intersection
SeqProfCod[a,d]	0.82	0.88	0.84	0.68	0.77	0.59	Dec05[d]
SNPs&GO[b,e]	0.87	0.90	0.88	0.72	0.77	0.65	Dec05
SeqProfCod[a,d]	0.74	0.65	0.78	0.83	0.72	0.48	Dec06[d]
SNPs&GO[b,e]	0.87	0.86	0.82	0.88	0.90	0.73	Dec06

[a]The method has been previously described [Capriotti et al., 2008] and includes on top of SeqProf information derived by computing the evolutionary selective pressure at the codon level.
[b]This article.
[c]The predictor with the same input as the one depicted in Figure 1, and one extra node to include selective pressure [Capriotti et al., 2008]. Intersection: the fraction of the all data set (28%) for which selective pressure parameters were available at http://sgu.bioinfo.cipf.es/services/Omidios (see text).
[d]Results previously published and obtained adopting a crossvalidation procedure on smaller sets than the one described in this artricle [Capriotti et al., 2008].
[e]Obtained from http://snps-and-go.biocomp.unibo.it.

Table 3. Bench Marking of SNPs&GO with Other Predictive Methods

Method	Q2	P[D]	Q[D]	P[N]	Q[N]	C	PM (%)
PolyPhen[a]	0.71	0.76	0.75	0.63	0.64	0.39	58
SIFT[b]	0.76	0.75	0.76	0.77	0.75	0.52	93
PANTHER[c]	0.74	0.77	0.73	0.71	0.76	0.48	76
Eremorph[d]	0.74	0.83	0.64	0.68	0.85	0.50	82
HybridMeth[e]	0.74	0.74	0.70	0.74	0.77	0.47	100
SNPs&GO	0.82	0.83	0.78	0.80	0.85	0.63	100

[a]Downloaded from the Web server http://genetics.bwh.harvard.edu/pph/data/index.html.
[b]Downloaded from http://blocks.fhcrc.org/sift/SIFT.html.
[c]Downloaded from http://www.pantherdb.org/downloads/index.jsp.
[d]Our previous implementation [Capriotti et al., 2006]. Performances are computed by training/testing on the SP_human data set (see Materials and Methods). Only HybridMeth and SNPs&GO are scored with a real crossvalidation procedure, because all the others were trained on sets that were possibly homologous to ours. PM is the percentage of predicted mutations. For the definition of the scoring indexes see the Materials and Methods section.
[e]Downloaded from the web server http://discovery.swmed.edu/eremorph/. Downloadable (a and c) predictors were run locally; otherwise, data are retrieved via Web server queries (a and d).

given type of diseases. The disease classification is derived from a previous work [Goh et al., 2007]. Apparently SNPs&PGO is better performing in the prediction of some types of diseases: respiratory, skeletal, cardiovascular, connective_tissue, bone, dermatological, metabolic, multiple, ear_nose_throat. For mutations linked to these diseases, the general level of accuracy is higher than 78% (compare with the overall SNPS&GO performance in Table 1). Somewhat lower levels of coverage are obtained for families of disease like: ophthamological, gastrointestinal, neurological, renal, hematological, muscular, cancer, and immunological. The results seem to be uncorrelated with the number of mutations in the different subsets.

The table list also the most significant GO terms (with LGO values ≥1) corresponding to each disease type that were highlighted with our analysis. The listed GO terms are the top scoring for each of the two GO functional branches (biological process and molecular function). Apparently subcellular localization (the third seed of functional terms in GO) is not significant. The biological processes well correlate with the disease type and the transcription factor activity is the most shared molecular function. In any case, these results indicate that even under the worst conditions the value of the type-specific index is not very far from that evaluated on the overall data set of mutations.

Discussion

The enormous number of human SNPs available in the databases and the ongoing jump in data volumes caused by ultrahigh-throughput sequencing (deep sequencing) poses the question of relating mutations to diseases. In this article, we propose a new SVM-based method that, starting from the protein sequence, uses different pieces of information, including that derived from the GO annotation of the protein to predict if a given mutation can be classified disease-related or not. This step can be important in prioritisation studies when SNPs need to be annotated for the selection of candidates genes in relation to a disease phenotype.

For the first time we present a GO-integrated predictor tested and trained with a stringent crossvalidation procedure. SNPs&GO was trained on a set of more than 33,000 annotated mutations in proteins, much larger than available before, and tested with a crossvalidation procedure over sets in which similar proteins were kept in the same data set also for the calculation of the LGO score, as derived from the GO database. With increasing complexity of information, the performance is enhanced, suggesting that in addition to the sequence profile, the LGO data derived from GO annotation improves our ability to discriminate neutral and disease-related SNPs. This adds to a previous analysis [Capriotti et al., 2008] in which we exploited the relevance of selective pressure as computed at the DNA level. However, at present, a wide-scale computation of parameters indicative of selective pressure is not feasible. We therefore confine the analysis at the protein level, describe the added value of GO terms to the prediction process, and compare with other available predictors in the Web. The finding that the level of performance increases with increasing information added to the input corroborates the notion that support vector machines can capture all the correlations existing in complementary knowledge. The benchmark that we

Table 4. The Prediction Accuracy Sorted by Disease Type and Most Significant GO Terms

Disease type	Q[D]	N_mut	GO terms[a]
Metabolic	0.81	3436	GO0006082 (organic acid metabolic process, BP) GO0048037 (cofactor binding, MF)
Neurological	0.74	1822	GO0007399 (nervous system development, BP) GO0022857 (transmembrane transporter activity, MF)
Cancer	0.73	1574	GO0048519 (negative regulation of biological process, BP) GO0043566 (structure-specific DNA binding, MF)
Hematological	0.71	1442	GO0009611 (response to wounding, BP) GO0019842 (vitamin binding, MF)
Ophthamological	0.76	1166	GO0050953 (sensory perception of light stimulus, BP) GO0003700 (transcription factor activity, MF)
Multiple	0.79	1117	GO0050954 (sensory perception of mechanical stimulus, BP) GO0003700 (transcription factor activity, MF)
Cardiovascular	0.88	706	GO0006936 (muscle contraction, BP) GO0008307 (structural constituent of muscle, MF)
Dermatological	0.86	667	GO0007398 (ectoderm development, BP) GO0005198 (structural molecule activity, MF)
Endocrine	0.77	648	GO0045941 (positive regulation of transcription, BP) GO0003700 (transcription factor activity, MF)
Renal	0.74	619	GO0007588 (excretion, BP) GO0022832 (voltage-gated channel activity, MF)
Muscular	0.77	508	GO0006936 (muscle contraction, BP) GO0005198 (structural molecule activity, MF)
Immunological	0.65	493	GO0009605 (response to external stimulus, BP) GO0042802 (identical protein binding, MF)
Connective_tissue	0.87	438	GO0030198 (extracellular matrix organization and biogenesis, BP) GO0005201 (extracellular matrix structural constituent, MF)
Bone	0.86	371	GO0001501 (skeletal development, BP) GO0005201 (extracellular matrix structural constituent, MF)
Skeletal	0.90	346	GO0001501 (skeletal development, BP) GO0019199 (transmembrane receptor protein kinase activity, MF)
Developmental	0.69	325	GO0048513 (organ development, BP) GO0003700 (transcription factor activity, MF)
Ear_Nose_Throat	0.78	199	GO0007605 (sensory perception of sound, BP) GO0005516 (calmodulin binding, MF)
Respiratory	0.93	162	GO0007585 (respiratory gaseous exchange, BP) GO0043492 (ATPase activity, coupled to movement of substances, MF)
Gastrointestinal	0.76	159	GO0048856 (anatomical structure development, BP) GO0022804 (active transmembrane transporter activity, MF)
Nutritional	0.33	3	GO0045600 (positive regulation of fat cell differentiation, BP) GO0003707 (steroid hormone receptor activity, MF)
Psychiatric	0.67	3	GO0055070 (copper ion homeostasis BP) GO0008017 (microtubule binding, MF)

For each disease type, as defined in ref [Goh et al., 2007], it is reported the coverage of predictions of the disease-related mutation (Q[D]). Q[D] is the number of mutations correctly predicted as disease-related over the total number of disease mutations belonging to a given disease class. N_mut is the total number of mutation belonging to each disease type.
[a]The most significant Go terms (with the highest LGO values for the disease type, ≥ 1) are listed; MF, molecular function, BP Biological Process.

performed in-house indicates that presently SNPs&GO is one of the best-scoring classifiers available for predicting whether a mutation at the protein level is or is not disease related.

The prediction accuracy was also sorted by disease type, and we found (Table 4) that it is rather independent of the morbidity. Furthermore, the most significant GO terms highlighted by our procedure well correlates with the disease type. These findings, all together, corroborate the view that our classifier on the basis of the input complexity that include the sequence function, is general enough to well discriminate whether a SNP occurring in a protein is disease related or not, rather independently of the disease type. The predictor is available for testing at http://snps-and-go.biocomp.unibo.it.

Acknowledgments

We thank MIUR for the PNR 2003 project (FIRB art.8) termed LIBI-Laboratorio Internazionale di BioInformatica delivered to R. Casadio. This work was also supported by the Biosapiens Network of Excellence project (a grant of the European Union's VI Framework Programme) and by the Strategic Project 2006 of the University of Bologna "Cross allergens in pollen and fruits: modulation by climate changes of their allergenic potential." R. Calabrese is supported by a FIRB 2003-LIBI grant. E.C. acknowledges support from a Marie Curie International Reintegration Grant (FP6-039722) delivered to Marc A. Marti-Renom.

References

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402.

Arbiza L, Duchi S, Montaner D, Burguet J, Pantoja-Uceda D, Pineda-Lucena A, Dopazo J, Dopazo H. 2006. Selective pressures at a codon-level predict deleterious mutations in human disease genes. J Mol Biol 358:1390–1404.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25:25–29.

Baldi P, Brunak S, Chauvin Y, Andersen CA, Nielsen H. 2000. Assessing the accuracy of prediction algorithms for classification: an overview. Bioinformatics 16:412–424.

Bao L, Cui Y. 2005a. Prediction of the phenotypic effects of non-synonymous single nucleotide polymorphisms using structural and evolutionary information. Bioinformatics 21:2185–2190.

Bao L, Zhou M, Cui Y. 2005b. nsSNPAnalyzer: identifying disease-associated nonsynonymous single nucleotide polymorphisms. Nucleic Acids Res 33:W480–W482.

Barbujani G, Goldstein DB. 2004. Africans and Asians abroad: genetic diversity in Europe. Annu Rev Genomics Hum Genet 5:119–150.

Bell J. 2004. Predicting disease using genomics. Nature 429:453–456.

Bishop CM. 2006. Pattern recognition and machine learning. Berlin: Springer.

Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O'Donovan C, Phan I, Pilbout S, Schneider M. 2003. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res 31:365–370.

Boyle EI, Weng S, Gollub J, Jin H, Botstein D, Cherry JM, Sherlock G. 2004. GO: TermFinder—open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. Bioinformatics 20:3710–3715.

Bromberg Y, Rost B. 2007. SNAP: predict effect of non-synonymous polymorphisms on function. Nucleic Acids Res 35:3823–3835.

Capriotti E, Arbiza L, Casadio R, Dopazo J, Dopazo H, Marti-Renom MA. 2008a. Use of estimated evolutionary strength at the codon level improves the prediction of disease-related protein mutations in humans. Hum Mutat 29:198–204.

Capriotti E, Calabrese R, Casadio R. 2006. Predicting the insurgence of human genetic diseases associated to single point protein mutations with support vector machines and evolutionary information. Bioinformatics 22:2729–2734.

Care MA, Needham CJ, Bulpitt AJ, Westhead DR. 2007. Deleterious SNP prediction: be mindful of your training data! Bioinformatics 23:664–672.

Cargill M, Altshuler D, Ireland J, Sklar P, Ardlie K, Patil N, Shaw N, LaneCR, Lim EP, Kalyanaraman N, Nemesh J, Ziaugra L, Friedland L, Rolfe A, Warrington J, Lipshutz R, Daley GQ, Lander ES. 1999. Characterization of single-nucleotide polymorphisms in coding regions of human genes. Nat Genet 22:231–238.

Chang CC, Lin CJ. 2001. Training nu-support vector classifiers: theory and algorithms. Neural Comput 13:2119–2147.

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

Cheng TM, Lu YE, Vendruscolo M, Lió P, Blundell TL. 2008 . Prediction by graph theoretic measures of structural effects in proteins arising from non-synonymous single nucleotide polymorphisms. PLoS Comput Biol 4:e1000135.

Collins FS, Guyer MS, Charkravarti A. 1997. Variations on a theme: cataloging human DNA sequence variation. Science 278:1580–1581.

Cotton RG, Auerbach AD, Axton M, Barash CI, Berkovic SF, Brookes AJ, Burn J, Cutting G, den Dunnen JT, Flicek P, Freimer N, Greenblatt MS, Howard HJ, Katz M, Macrae FA, Maglott D, Möslein G, Povey S, Ramesar RS, Richards CS, Seminara D, Smith TD, Sobrido MJ, Solbakk JH, Tanzi RE, Tavtigian SV, Taylor GR, Utsunomiya J, Watson M. 2008. GENETICS. The human variome project. Science 322:861–862.

Dayhoff MO, Schwartz RM, Orcutt BC. 1978. A model of evolutionary change in proteins. Atlas Protein Sequence Struct 5:345–352.

Dobson RJ, Munroe PB, Caulfield MJ, Saqi MA. 2006. Predicting deleterious nsSNPs: an analysis of sequence and structural attributes. BMC Bioinformatics 7:217–226.

Edmonds CA, Lillie AS, Cavalli-Sforza LL. 2004. Mutations arising in the wave front of an expanding population. Proc Natl Acad Sci USA 101:975–979.

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 315:771–786.

Ferrer-Costa C, Orozco M, de la Cruz X. 2004. Sequence-based prediction of pathological mutations. Proteins 57:811–819.

Ferrer-Costa C, Orozco M, de la Cruz X. 2005. Use of bioinformatics tools for the annotation of disease-associated mutations in animal models. Proteins 61:878–887.

Goh KI, Cusick ME, Valle D, Childs B, Vidal M, Barabási AL. 2007. The human disease network. Proc Natl Acad Sci USA 104:8685–8690.

Goldstein DB, Cavalleri GL. 2005. Genomics: understanding human diversity. Nature 437:1241–1242.

Henikoff S, Henikoff JG. 1992. Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci USA 89:10915–10919.

Kaminker JS, Zhang Y, Waugh A, Haverty PM, Peters B, Sebisanovic D, Stinson J, Forrest WF, Bazan JF, Seshagiri S, Zhang Z. 2007a. Distinguishing cancer-associated missense mutations from common polymorphisms. Cancer Res 67:465–473.

Kaminker JS, Zhang Y, Watanabe C, Zhang Z. 2007b. CanPredict: a computational tool for predicting cancer-associated missense mutations. Nucleic Acids Res 35:W595–W598.

Krishnan VG, Westhead DR. 2003. A comparative study of machine-learning methods to predict the effects of single nucleotide polymorphisms on protein function. Bioinformatics 19:2199–2209.

Kulkarni V, Errami M, Barber R, Garner HR. 2008. Exhaustive prediction of disease susceptibility to coding base changes in the human genome. BMC Bioinformatics. 9(Suppl 9):S3.

Ng PC, Henikoff S. 2002. Accounting for human polymorphisms predicted to affect protein function. Genome Res 12:436–446.

Ng PC, Henikoff S. 2003. SIFT: Predicting amino acid changes that affect protein function. Nucleic Acids Res 31:3812–3814.

Pei J, Grishin NV. 2001. AL2CO: calculation of positional conservation in a protein sequence alignment. Bioinformatics 17:700–712.

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

Rish N, Merikangas K. 1996. The future of genetic studies of complex human diseases. Science 273:1516–1517.

Riva A, Kohane IS. 2002. SNPper: retrieval and analysis of human SNPs. Bioinformatics 18:1681–1685.

Robert J, Morvan VL, Smith D, Pourquier P, Bonnet J. 2005. Predicting drug response and toxicity based on gene polymorphisms. Crit Rev Oncol Hematol 54:171–196.

Schwarz DF, Hädicke O, Erdmann J, Ziegler A, Bayer D, Möller S. 2008. SNPtoGO: characterizing SNPs by enriched GO terms. Bioinformatics 24:146–148.

Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K. 2001. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res 29:308–311.

Stitziel NO, Binkowski TA, Tseng YY, Kasif S, Liang J. 2004. topoSNP: a topographic database of non-synonymuous single nucleotide polymorphism with and without known disease association. Nucleic Acids Res 32:D520–D522.

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

Tavtigian SV, Greenblatt MS, Lesueur F, Byrnes GB, IARC Unclassified Genetic Variants Working Group. 2008. In silico analysis of missense substitutions using sequence-alignment based methods. Hum Mutat 29:1327–1336.

Thomas PD, Campbell MJ, Kejariwal A, Mi H, Karlak B, Daverman R, Diemer K, Muruganujan A, Narechania A. 2003. PANTHER: a library of protein families and subfamilies indexed by function. Genome Res 13:2129–2141.

Thomas PD, Kejariwal A. 2004. Coding single-nucleotide polymorphisms associated with complex vs. Mendelian disease: evolutionary evidence for differences in molecular effects. Proc Natl Acad Sci USA 101:15398–15403.

Tian J, Wu N, Guo X, Guo J, Zhang J, Fan Y. 2007. Predicting the phenotypic effects of non-synonymous single nucleotide polymorphisms based on support vector machines. BMC Bioinformatics 8:450.

Wang DG, Fan JB, Siao CJ, Berno A, Young P, Sapolsky R, Ghandour G, Perkins N, Winchester E, Spencer J, Kruglyak L, Stein L, Hsie L, Topa-loglou T, Hubbell E, Robinson E, Mittmann M, Morris MS, Shen N, Kil-burn D, Rioux J, Nusbaum C, Rozen S, Hudson TJ, Lipshutz R, Chee M, Lander ES. 1998. Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. Science 280:1077–1082.

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

Worth CL, Bickerton GR, Schreyer A, Forman JR, Cheng TM, Lee S, Gong S, Burke DF, Blundell TL. 2007. A structural bioinformatics approach to the analysis of nonsynonymous single nucleotide polymorphisms (nsSNPs) and their relation to disease. J Bioinform Comput Biol 5:1297–1318.

Yip YL, Scheib H, Diemand AV, Gattiker A, Famiglietti LM, Gasteiger E, Bairoch A. 2004. The Swiss-Prot variant page and the ModSNP database: a resource for sequence and structure information on human protein variants. Hum Mutat 23:464–470.

Yue P, Li Z, Moult J. 2005. Loss of protein structure stability as a major causative factor in monogenic disease. J Mol Biol 353:459–473.

Yue P, Moult J. 2006. Identification and analysis of deleterious human SNPs. J Mol Biol 356:1263–1274.