HUMAN MUTATION 29(1), 198–204, 2008

METHODS

Use of Estimated Evolutionary Strength at the Codon Level Improves the Prediction of Disease-Related Protein Mutations in Humans

Emidio Capriotti,[1] Leonardo Arbiza,[2] Rita Casadio,[4] Joaquín Dopazo,[3] Hernán Dopazo,[2]* and Marc A. Marti-Renom[1]*

[1]Structural Genomics Unit, Centro de Investigación Príncipe Felipe (CIPF), Valencia, Spain; [2]Pharmacogenomics and Comparative Genomics Unit, Centro de Investigación Príncipe Felipe (CIPF), Valencia, Spain; [3]Functional Genomics Unit, Bioinformatics Department, Centro de Investigación Príncipe Felipe (CIPF), Valencia, Spain; [4]Laboratory of Biocomputing, CIRB/Department of Biology, University of Bologna, Bologna, Italy

Communicated by David N. Cooper

Predicting the functional impact of protein variation is one of the most challenging problems in bioinformatics. A rapidly growing number of genome-scale studies provide large amounts of experimental data, allowing the application of rigorous statistical approaches for predicting whether a given single point mutation has an impact on human health. Up until now, existing methods have limited their source data to either protein or gene information. Novel in this work, we take advantage of both and focus on protein evolutionary information by using estimated selective pressures at the codon level. Here we introduce a new method (SeqProfCod) to predict the likelihood that a given protein variant is associated with human disease or not. Our method relies on a support vector machine (SVM) classifier trained using three sources of information: protein sequence, multiple protein sequence alignments, and the estimation of selective pressure at the codon level. SeqProfCod has been benchmarked with a large dataset of 8,987 single point mutations from 1,434 human proteins from SWISS-PROT. It achieves 82% overall accuracy and a correlation coefficient of 0.59, indicating that the estimation of the selective pressure helps in predicting the functional impact of single-point mutations. Moreover, this study demonstrates the synergic effect 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 the codon level. The results of large-scale application of SeqProfCod over all annotated point mutations in SWISS-PROT (available for download at http://sgu.bioinfo.cipf.es/services/Omidios/; last accessed: 24 August 2007), could be used to support clinical studies. Hum Mutat 29(1), 198–204, 2008. **© 2007 Wiley-Liss, Inc.**

KEY WORDS: SNP; nsSNP; disease; sequence profile; evolutionary strength; bioinformatics

INTRODUCTION

Studies characterizing the relationship between protein variants and human disease have grown rapidly over the past years, in part due to genomic-scale sequencing efforts [Krawczak et al., 2000; Sherry et al., 2001; Stenson et al., 2003]. For example, it is now known that single nucleotide polymorphisms (SNPs) constitute about the 90% of human protein sequence variability [Collins et al., 1998]. Synonymous and nonsynonymous SNPs (nsSNPs) may occur every ~350 bp in coding regions [Cargill et al., 1999] and about 50% of nsSNPs may be associated to pathologies of genetic origin. Therefore, predicting which nsSNPs are responsible for human disease is one of the major challenges in bioinformatics.

Recently, different methods have been developed for predicting the effect of single point mutations in humans [Arbiza et al., 2006; Bao and Cui, 2005; Bao et al., 2005; Capriotti et al., 2006; Chan et al., 2007; Karchin et al., 2005a; Ng and Henikoff, 2003; Ramensky et al., 2002; Santibanez Koref et al., 2003; Thomas et al., 2003b; Yue and Moult, 2006]. In spite of the effort, however,

Received 30 May 2007; accepted revised manuscript 17 July 2007.

*Correspondence to: Marc A. Marti-Renom and Hernán Dopazo, Bioinformatics Department, Centro de Investigación Príncipe Felipe (CIPF). Av. Autopista del Saler, 16, 46013 Valencia, Spain. E-mail for Marc A. Marti-Renom: mmarti@cipf.es; E-mail for Hernán Dopazo hdopazo@cipf.es

Grant sponsor: Ministerio dell' Universitá e della Ricerca, Italy; Grant: Fondo per gli Investimenti della Ricerca di Base 2003 LIBI-International Laboratory of Bioinformatics; Grant sponsor: Marie Curie International Reintegration Grant; Grant number: FP6-039722; Grant sponsor: Generalitat Valenciana; Grant numbers: GV/2007/065 and GV06/080; Grant sponsor: Ministerio de Educación y Ciencia, Spain; Grant number: BFU2006-15413-C02-02/BMC; Grant sponsor: European Union, (EU) Network of Excellence BIOSAPIENS: Grant number: LSHG-CT-2003-503265.

DOI 10.1002/humu.20628

Published online 12 October 2007 in Wiley InterScience (www.interscience.wiley.com).

the identification of disease-associated human nsSNPs remains a difficult task and a satisfactory solution of general applicability is yet unavailable. Routinely, two different types of information have been used to address the problem. On the one hand, most of the previously published works focus their attention toward information from protein sequences and/or their homologs. For example, the putative effect of mutations has been predicted by adopting protein multiple sequence alignments [Capriotti et al., 2006; Ng and Henikoff, 2003; Ramensky et al., 2002; Thomas et al., 2003b], protein structures [Anishetty et al., 2006; Karchin et al., 2005a; Terp et al., 2002; Yue and Moult, 2006], or both [Bao and Cui, 2005; Ramensky et al., 2002]. On the other hand, recent studies have used codon-based information within a phylogenetic framework to assess the degree of association between a point mutation and its possible pathogenic effects [Arbiza et al., 2006; Santibanez Koref et al., 2003]. A common evolutionary approach to determine selective pressure acting at a molecular level involves the estimation of the ratio of nonsynonymous and synonymous rates of substitution per site ($\omega = dN/dS$) [Yang, 2003]. Based on 43 genes, we have recently hypothesized that residues evolving under markedly strong selective pressures ($\omega < 0.1$) are significantly ($P < 0.01$) associated with human disease [Arbiza et al., 2006].

Here we propose that: 1) the estimation of the selective pressure can be used for large-scale functional annotation of nsSNPs; and 2) the combination of protein sequence/profile-based information with codon-based information increases the accuracy of disease prediction. We begin by describing the benchmarking datasets, the protocol for building multiple sequence alignments, the methods to estimate selective pressures, the building of SVMs, and the accuracy measures (see Materials and Methods). We then assess the results of our new classifiers and discuss the implications for predicting the likeliness of a point mutation to be associated or not with human disease (see Results and Discussion). Finally, we outline the main conclusions of the present work.

MATERIALS AND METHODS

Datasets

The selection of the training and testing datasets can affect the accuracy of predicting deleterious and neutral effects of protein variants. Our datasets were extracted from the SWISS-PROT subset of the UniProt database [Bairoch et al., 2005], which was recently described as the best dataset for training [Care et al., 2007]. SWISS-PROT classifies protein variants as disease related (i.e., with pathological effects), polymorphism (i.e., with no effect on human health), or unclassified. For this study, two different sets were obtained: SP-Dec05, derived from the SWISS-PROT release 48 (Dec 2005) and SP-Dec06, which included only mutations from protein sequences deposited in SWISS-PROT from January to November 2006 (release 51). The SP-Dec06 dataset was only used for testing the robustness of our method. Five different filters were then applied to each of the two datasets to: 1) remove sequences from organisms other than *Homo sapiens*; 2) remove all protein variants that were unclassified; 3) remove all variants other than single point mutations; 4) remove all variants for which the substitution rate could not be calculated using the Ensembl database; and 5) remove all variants with less than 10 aligned sequences in its multiple sequence alignment. The SP-Dec05 and SP-Dec06 datasets included a total of 8,987 and 2,008 protein variants, respectively (Table 1). The complete datasets are available for download at http://sgu.bioinfo.cipf.es/datasets.

TABLE 1. nsSNP Datasets Used in This Work

	Protein variants	Disease	Polymorphism	Sequences
SP-Dec05	8,987	6,220	2,767	1,434
SP-Dec06	2,008	804	1,204	720

Protein Sequence Profiles

For each protein in the SP-Dec05 and SP-Dec06 datasets, a sequence profile was built by collecting homologous sequences from a nonredundant database at 95% sequence identity (nr95, release September 2006) using the BLAST algorithm with an inclusion e-value threshold of 10^{-9} and all other parameters set to their default values [Altschul et al., 1990]. This nr95 database was built from the NCBI nonredundant sequence database by running the *cd-hit* algorithm with 95% sequence identity cutoff [Li and Godzik, 2006].

Selective Pressures

Orthologous sequences in eight mammalian species for 2,466 genes were retrieved from the Ensembl-Compara database (version v.42) of the Ensembl Database Project [Hubbard et al., 2005]. The genomes were as follows: 1) human (*Homo sapiens*, v.42_36d); 2) chimpanzee (*Pan troglodytes*, v.42_21a); 3) macaque (*Macaca mulatta*, v.42_10b); 4) mouse (*Mus musculus*, v.42_36c); 5) rat (*Ratus norvegicus*, v.42_34l); 6) dog (*Canis familiaris*, v.42_2); 7) bull (*Bos taurus*, v.42_2e); and 8) opossum (*Monodelphis domestica*, v.42_3b). If multiple orthologous relationships were annotated in the database, the relationship showing the shortest distance with the Ensembl tree was selected. In this work, the phylogenetic relationships between the eight species were established according the following topology: (((((1,2),3),(4,5)),(6,7)),8) [Springer et al., 2004].

Translated protein sequences based on the largest transcript of each gene in the different species were aligned using Muscle v3.6 with a maximum of 10,000 iterations or 5 hr running time and all other parameters set to default values [Edgar, 2004]. Gapped positions in the resulting multiple protein sequence alignments were removed before mapping back to the DNA sequence and estimating the selective pressure both for each lineage and at each position through different approaches. Estimates of selective pressure (ω) for each lineage were obtained using a free branch model and the F3x4 codon frequency option as implemented in the he *codeml* program of the PAML (version 3.14a) package [Yang, 1997]. Both the nonsynonymous rate of substitution per nonsynonymous site (dN) and the synonymous rate of substitution per synonymous site (dS) used in the estimation of lineage-wise selective pressure ($\omega = dN/dS$) were taken into account. Two different likelihood site-models (M2a and M8) from *codeml* were additionally used to estimate the selective pressure (ω) at given codon sites. All parameters for the site models were set to their default values except for the "runmode" and "model" parameters that were set to zero, the codon frequency model was set to F3x4, and the "Nssites" parameter was set to 2 and 8 for the M2a and M8 models, respectively. Empirical Bayes analysis [Nielsen et al., 1998; Yang et al., 2005] was used to calculate the posterior probability of each site belonging to each class in the models M2a and M8. The analysis gave us the posterior ω values, which represent the strength of natural selection acting on each site. These posterior ω values denoting purifying selection, positive selection, or neutral evolution on sites, were the values used for all proteins evaluated under the codon-based likelihood site models

M2a and M8 (i.e., ωM2 and ωM8, respectively). For each protein in the SP-Dec05 and SP-Dec06 datasets, the model with the best likelihood was selected. Finally, the Kolmogorov Smirnov (K-S) test, implemented in the R-statistical package [Ihaka and Gentelman, 1996], was used to evaluate the statistical significance of the difference between ω value distributions associated to disease and polymorphism.

Support Vector Machine Classifiers

Support vector machines (SVMs) are universal classifiers that learn a variety of data distributions from training samples, and as such, are applicable to classification and regression tasks [Vapnik, 1995]. SVMs have previously been used for predicting the phenotypic effect of a point mutation in a protein [Bao and Cui, 2005; Bao et al., 2005; Capriotti et al., 2005a; Chan et al., 2007; Karchin et al., 2005a, 2005b; Yue and Moult, 2006]. Here we rely on our previous work [Capriotti et al., 2006, 2005a, 2005b] to extend the implementation of sequence-based and profile-based SVMs to include codon-based estimation of selective pressures at each position of the target sequences. Our SVMs were trained with a radial basis function standard kernel implemented in the LIBSVM package [Chang and Lin, 2001]. The *grid* program from the same package was used to search for the optimal "C" and "G" parameters (i.e., here set to 0.5 and 3.05×10^{-5}, respectively). In this work, we have developed four types of SVMs depending on the input information: 1) sequence-based SVMs that encode for protein sequence information (Seq); 2) sequence- and profile-based SVMs that encode for sequence and profile information (SeqProf); 3) sequence- and codon-based SVMs that encode both for protein sequence and selective pressure (SeqCod); and 4) combined SVMs that use all the information available considering the protein sequence, the sequence profile and selective pressure at the codon level (SeqProfCod).

First, the Seq SVM was trained using only sequence information encoded in a vector of 40 elements or features [Capriotti et al., 2006, 2005a]. The first 20 elements encode for the change in amino acid type at the mutation site and the second 20 elements encode for the frequency of amino acid types in a window of 18 residues around this position.

Second, SeqProf was trained by adding the information derived from a multiple sequence alignment of the target sequence and its close homologous sequences to the Seq classifier. Two new features were added to the previous 40-element vector: the ratio of the mutated residue vs. the wild-type residue in the sequence profile and the number of aligned sequences in the mutated position encoded for the profile information.

Third, the SeqCod SVM was trained using sequence- and codon-based information encoded in a vector of 43 elements of which the sequence information is encoded in the first 40 elements and codon-based information is encoded in the last three elements corresponding to the codon-based ω and lineage-based estimates of dN and dS.

Fourth, SeqProfCod combined protein sequence and profile information with selective pressures in a 45 features vector (i.e., 40 features describing the sequence-based information, two features describing the profile-based information, and three features describing the codon-based information).

Protein Profile-Based Methods

Two widely used methods based on protein profile information were also tested against the combined SeqProfCod classifier. First, the SIFT program [Ng and Henikoff, 2002], which is based on the premise that important amino acids are conserved within a protein family, was downloaded from http://blocks.fhcrc.org/sift/SIFT.html [Ng and Henikoff, 2003] and run locally with all its parameters set to their default values. Second, the PANTHER program [Thomas et al., 2003a], which uses a profile-based Hidden Markov Model to rank polymorphisms according to their likelihood of affecting protein function, was downloaded from www.pantherdb.org and run locally against the PANTHER_6.1 library with all its parameters set to their default values.

Accuracy Measures

The resulting classifiers were tested for their accuracy for predicting the phenotypic effects of point mutations using a cross-validation procedure over the SP-Dec05 dataset. Their accuracy was calculated by a 20-fold cross-validation procedure, which randomly replicated neutral polymorphisms mutation data to balance its proportion to disease-related mutations in the training sets. Furthermore, to prevent overtraining and overestimating of the results due to the redundancy of sequences, all the proteins in the datasets were clustered according to their sequence similarity using the *blastclust* program with its parameters set to their default values [Altschul et al., 1997]. No two proteins from the same cluster were allowed to be part of the same training and testing datasets.

The overall accuracy ($Q2$) of a binary classifier is calculated as:

$$Q_2 = \frac{TP + TN}{N} \tag{1}$$

where TP an TN are the true positive and negative predicted mutations and N is the total number of mutations. The correlation coefficient C is:

$$C = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FN) \times (TP + FP) \times (TN + FP) \times (TN + FN)}} \tag{2}$$

where FP and FN are the false positive and negative predicted mutations.

The accuracy $Q(s)$ for each class s and its complementary (i.e., disease and polymorphism) is:

$$Q(s) = \frac{T(s)}{T(s) + F(\bar{s})} \tag{3}$$

A reliability score ($RI(s)$) to each prediction is:

$$RI(s) = 10 \times abs(O(s) - t) \times w(s) \tag{4}$$

where $O(s)$ is the probability assigned by the classifier to the class s, t is a threshold, and $w(s)$ is the weight of the set relative to the class s. In this work, t and $w(s)$ were set to 0.5.

Other standard scoring measures [Baldi et al., 2000], including the area under the ROC curve (AUC) and the true-positive rate (TPR) and false-positive rate (FPR) are also used to assess the accuracy of the benchmarked methods. TPR and FPR are:

$$TPR = \frac{TP}{TP + FN} \quad FPR = \frac{FP}{FP + TN} \tag{5}$$

RESULTS AND DISCUSSION

Selective Pressures, Human Disease, and Polymorphism

Natural selection works in proportion to the number of deleterious mutations occurring in the population [Kimura, 1983]. On the one hand, mutations on functionally relevant residues are expected to show high selective constraints. On the

other hand, residues that are not associated with major functional roles of the protein may be changing under neutral evolution and consequently not necessarily found in association with disease. Accordingly, we hypothesized that nsSNPs from coding regions of the genome that affect human health may evolve more frequently under strong selective pressure (i.e., $\omega \leq 0.1$) [Arbiza et al., 2006]. Using a large-scale testing set, we have found a statistically significant association in humans between high selective pressures and disease, in contrast to low selective pressures and neutral polymorphic variants (Fig. 1). Disease-related protein variants and polymorphisms show significantly different distributions. The median ω values for disease-related and neutral polymorphisms variants are 0.068 and 0.14, respectively. Therefore, the disease-related median value is 0.072 lower than that for polymorphisms. This difference, although small, is very significant given a much larger distribution for ω values of polymorphisms (P-value of 2.2×10^{-16}). This result indicates that ω values smaller than 0.1 are more frequently associated to disease than to polymorphisms.

Protein Sequence and Profile-Based Classifiers

The Seq classifier, which solely includes information about the target sequence, results in an average overall accuracy (Q2) of 0.73, correctly predicting 72% of the disease associated mutations and 74% of the polymorphisms in the SP-Dec05 dataset (Table 2). The correlation coefficient (C) is 0.43 and the area under the curve (AUC) is 0.81 (Table 2). This classifier can be considered the base line reference to assess the increment in accuracy as evolutionary information is added during training. The accuracy of SeqProf for predicting disease-related mutations compared to the Seq classifier increases from 72% to 80% with a final correlation of 0.52. This increment is also reflected in a larger AUC of 0.85 and the percentage of true-positive rate (TPR), which increases 7% points with respect Seq at the false-positive rate (FPR) of 5% (Fig. 2). The SeqProf method results in a higher accuracy than using the ratio of mutated residue alone calculated from the sequence profile. This shows that the introduction of sequence information with profile-based information improves the quality of the predictions (Fig. 2A). Thus, the results presented here are in

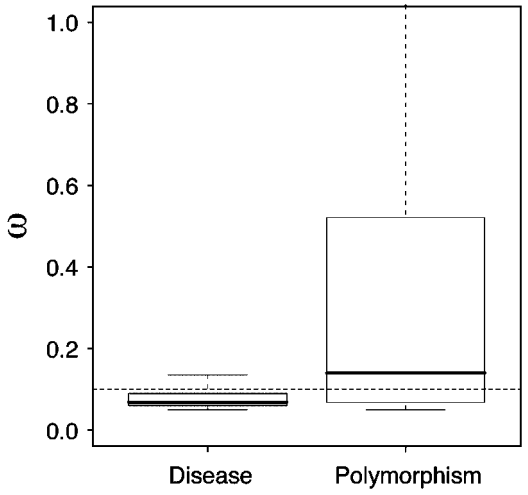

FIGURE 1. **ω distribution for disease and polymorphism protein variants in the SP-Dec05 dataset. The box-plot shows the upper and lower quartiles (box), the interquartile range (dashed vertical lines), and the median (horizontal bold line) values for disease-related and polymorphism protein variants (0.068 and 0.14, respectively). For visual inspection, a dashed horizontal line in gray indicates ω value of 0.1.**

TABLE 2. **Accuracy of the Classifiers Over the SP-Dec05 Dataset**

	Q2	Q (disease)	Q (neutral)	C	AUC
Seq	0.73	0.72	0.74	0.43	0.81
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

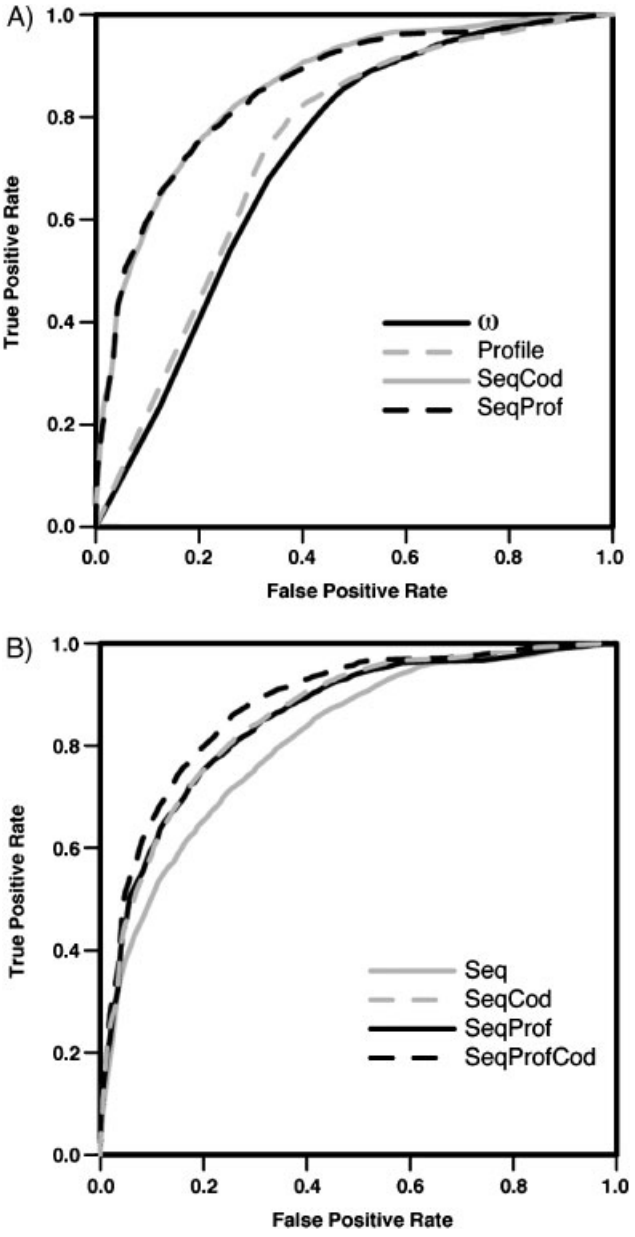

FIGURE 2. **Receiver operating characteristic (ROC) curves. The area under the ROC curve represents the probability of correct classification over the whole range of cutoffs. This area is usually taken to be an important index because it provides a single measure of overall accuracy that is not dependent upon a particular feature threshold. A: Comparison of individual input scores ω and ratio of mutated residue against SVM trained SeqCod and SeqProf. B: Comparison of our four SVM-based methods Seq, SeqCod, SeqProf, and SeqProfCod.**

agreement with our previous work that indicated the role of sequence environment and sequence profile in improving disease classification using SVM-based methods [Capriotti et al., 2006, 2005a].

Codon-Based SVMs

The SeqCod classifier, which takes into account the ω value for the mutated position and dN and dS for branches calculated with the PAML model, results in a Q2 of 0.79 (6% increase in accuracy with respect to Seq) and a AUC of 0.86. SeqCod correctly predicts 82% and 74% of the disease-related and polymorphism protein variants, respectively (Table 2). SeqCod predictions are also more accurate than simply using the ω measures with no combination with sequence and no training by the SVM (Fig. 2A). Including both protein-based and codon-based information further increases the accuracy of the classifier by several percentile points. For example, SeqProfCod results in a Q2 of 0.82, correctly predicting ~4% more protein variants than either SeqCod or SeqProf and ~10% more protein variants than Seq alone (Table 2). SeqProfCod results in an AUC of 0.88, the largest of the tested classifiers. In particular, SeqProfCod increases by ~12% points the value of TPR at 5% of FPR with respect to the sequence-based classifier (Fig. 2B). The overall accuracy and the correlation coefficient of the SeqProfCod classifier increase with the reliability index (RI) (Fig. 3). The RI, calculated for each prediction, is indicative of the accuracy of each of the predictions by SeqProfCod. At an RI of 3, SeqProfCod is able to predict an effect of a protein variant for the 89% of the dataset with a correlation coefficient of 0.69 (Fig. 3). Moreover, SeqProfCod reaches similar accuracy using the SP-Dec06 testing dataset showing a robust level of reliability predicting disease-related mutations (77%) even when this dataset is unbalanced toward neutral polymorphisms (Table 1).

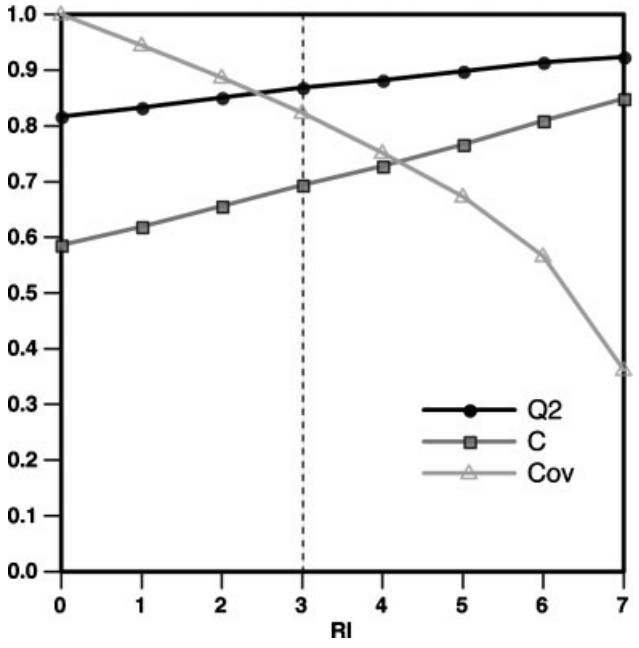

FIGURE 3. **Overall accuracy (Q2) and correlation (C) of SeqProfCod as a function of the reliability index (RI). Cov is the fraction of the SP-Dec05 dataset with RI values higher or equal to the given cutoff. The horizontal dashed line crosses the plot at the RI cutoff of 3, corresponding to Q2 of 0.87, C of 0.69, and Cov of 82%.**

Testing SeqProfCod Against SIFT and PANTHER

The results so far show that the accuracy of our classifiers increases with the inclusion of more detailed information about the evolution of the target sequences. These differences are clear when comparing Seq, which includes solely the sequence-based information, with SeqProfCod, which includes evolutionary information from protein- and codon-based multiple alignments. Compared to the SIFT and PANTHER programs, SeqProfCod results in similar or higher accuracies for both datasets (Tables 3 and 4). SIFT results in a Q2 of 0.71 over the SP-Dec05 and SP-Dec06 datasets. The accuracy of PANTHER is between 3% and 6% points higher than that of SIFT for both testing datasets (i.e., Q2 of 0.74 and 0.77, respectively). However, both methods, SIFT and PANTHER were unable to predict the effect of all protein variants in the testing sets. When a protein variant could not be aligned to either a block or a Hidden Markov Model, SIFT and PANTHER were unable to predict the likely outcome of the point mutations (i.e., 3% and 17% of the SP-Dec05 dataset and 4% and 23% of the SP-Dec06 dataset, respectively). SeqProfCod results in a higher accuracy than SIFT of ~11% in Q2 and 0.21 in C. Compared to PANTHER the increase in accuracy is smaller resulting in 8% higher Q2 and in 0.16 higher C (Table 3). However, the coverage of PANTHER is 18% and 23% smaller than that of SeqProfCod for the SP-Dec05 and SP-Dec06 datasets, respectively. Filtering our predictions and considering only those with RI larger than or equal to 3, SeqProfCod results in a Q2 equal to 80% and C of 0.59 over 78% of the mutations included in the SP-Dec06. Thus, at a similar coverage, our method results in higher accuracy than PANTHER.

Advantage of Combining ω and Sequence Profiles

SeqProfCod uses two main sources of information to predict the most probable outcome of a point mutation (i.e., multiple sequence alignment and estimation of evolutionary strength at the codon level). To discern the contribution of each of the factors as well as their synergy, we have studied the error rate (i.e., false predictions) at optimal cutoffs for both ω and the ratio of mutated residue. Minimum error rates are reached at ω of 0.12 and ratio of mutated residue of 0.07 for the SP-Dic05 dataset, which was then accordingly divided into four different subsets considering those two cutoffs (Table 5). On average, SeqProfCod results in more accurate predictions (~4% higher average accuracy with respect to SeqCod and SeqProf) when both ω and the ratio of mutated

TABLE 3. **Accuracy of SeqProfCod Compared to SIFT and PANTHER Over the SP-Dec05 Dataset**

	Q2	Q (disease)	Q (neutral)	C	Cov (%)
SeqProfCod	0.82	0.84	0.77	0.59	100
SIFT	0.71	0.72	0.69	0.38	97
PANTHER	0.74	0.75	0.72	0.43	83

TABLE 4. **Accuracy of SeqProfCod Compared to SIFT and PANTHER Over the SP-Dec06 Dataset**

	Q2	Q (disease)	Q (neutral)	C	Cov (%)
SeqProfCod	0.74	0.78	0.72	0.48	100
SIFT	0.71	0.70	0.72	0.42	96
PANTHER	0.77	0.71	0.81	0.52	77

TABLE 5. Performances of SeqCod, SeqProf, and SeqProfCod on Different Subsets of SP-Dec05

			SeqCod		SeqProf		SeqProfCod	
ω	Ratio mutated residue	Dataset%	Q2	C	Q2	C	Q2	C
≤0.12	≤0.07	56	0.84	0.30	0.84	0.30	0.88	0.32
>0.12	>0.07	14	0.81	0.34	0.80	0.32	0.81	0.31
≤0.12	>0.07	17	0.70	0.39	0.63	0.36	0.68	0.40
>0.12	≤0.07	13	0.68	0.41	0.73	0.43	0.74	0.49

residue are below their respective cutoffs. SeqProfCod also results in better accuracy for ~30% of nsSNPs in the SP-Dic05 dataset that result in ω and ratio of mutated residue with different tendency (i.e., ω >0.12 and ratio of mutated residue ≤0.07 or ω≤0.12 and mutation ratio >0.07). Therefore, the results indicate that the SeqProfCod is able to discern between the two features by selecting the most informative. For example, for ω >0.12 and ratio of mutated residue ≤0.07, SeqProfCod results in average accuracy closer to SeqProf than SeqCod. Similarly, with ω≤0.12 and ratio of mutated residue >0.07, SeqProfCod results in an average accuracy closer to SeqCod than SeqProf. Thus, SeqProfCod generally predicts disease or polymorphism if both cutoffs are below or above the optimal thresholds, respectively. Accordingly, SeqCod and SeqProf accuracies decrease when a disagreement between the two parameters occurs (Table 5). These results, as well as the two examples outlined next, show that SeqProfCod is able to capture the synergic effect of combining ω and sequence profiles.

Tumor Suppressor p53 and Iduronate 2-Sulfatase Precursor Genes

The p53 gene product (P53_HUMAN) acts as a tumor suppressor by inducing growth arrest or apoptosis, depending on the physiological circumstances and cell type of the tumor. Several mutations in the gene have been already associated with cancer. In particular, mutations P278A, P278T, R280K, and R283H are associated with higher risk of colon cancer [De Vries et al., 1996]. The ω calculated for the sequence positions 278, 280, and 283 was 0.13, 0.11, and 0.11, respectively. Such ω values are near or above the optimal cutoff thus yielding to polymorphism prediction when using SeqCod. However, those positions were completely conserved in the multiple sequence alignment, resulting in mutated residue ratios equal to zero, which made it possible to correctly predict those mutations as diseases associated by SeqProfCod. On average, SeqProfCod correctly predicts 81% of all 52 p53 mutations annotated in SWISS-PROT.

The Iduronate 2-sulfatase precursor gene (IDS_HUMAN) is responsible for the lysosomal degradation of proteoglycans. Two particular mutations in the IDS gene (S143F and P160R) have been associated with Hunter syndrome [Hopwood et al., 1993; Karsten et al., 1998]. The mutated residue ratio for S143F and P160R are 0.05 and 0.06, respectively, which are near the optimal cutoff and make it difficult to correctly predict them using only profile-based information (i.e., by SeqProf). However, the ω value for positions 143 and 160 was 0.08. Thus, using SeqProfCod correctly predicted the association of those two mutations with disease. The overall accuracy of SeqProfCod for all 106 IDS mutations annotated in SWISS-PROT was 92%.

Application to the CDKN2A, MLH1, MSH2, MECP2, and TYR Genes

In a recent article, Chan et al. [2007] compared four different computational methods for predicting the likely outcome of a point mutation in a limited number of human genes associated with inherited disorders (i.e., CDKN2A, a tumor suppressor; MSH2 and MLH1, responsible for the hereditary cancer syndrome; and MECP2 and TYR). The authors concluded that the tested methods were able to correctly predict the mutation effects for 73 to 82% of the 254 mutations dataset from the five genes. An interesting result from their work is that using a consensus approach and selecting only predictions in which most of the methods agreed, resulted in a significant increase of accuracy at the cost of decrease in coverage [Chan et al., 2007]. SeqProfCod could be applied to 47.6% of the 254 point mutations, resulting in an accuracy of 86.8%, which is comparable with the accuracy by the consensus approach proposed by Chan et al. [2007].

CONCLUSIONS

With this work we show that the estimation of selective pressures can be used for large-scale functional annotation of nsSNPs and that the combination of protein sequence/profile-based information with codon-based information increases the accuracy of disease prediction. Our initial hypothesis that codons with estimated ω values smaller than 0.1 were more strongly associated with disease mutations in human has been corroborated by the present analysis. Moreover, the combination of codon-based information with protein sequence- and profile-based information has yielded a new SVM classifier that results in higher accuracy than other tested methods. SeqProfCod may prove useful for annotating the effects of point mutations in genomic-scale predictions and could have an added value for clinical counseling in assessing the likely outcome of a SNP in a patient [Chan et al., 2007]. Moreover, SeqProfCod predictions are automatically associated with a reliability index, which makes them more useful for clinical decision-making. Although the absolute gains in terms of Q2 appear to be small, the benefits could telescope in a large-scale application such as predicting the effects of the ~57,000 nsSNPs annotated in dbSNP [Sherry et al., 2001]. By using SeqProfCod, we could expect to correctly predict the likely effect of a mutation for ~2,000 more nsSNPs than without using the codon-based information. An initial application at a genomic scale of our new classifier provides us a priori assessment of the phenotypic effect of nsSNPs in the human genome.

ACKNOWLEDGMENTS

M.A.M.R. and E.C. acknowledge support from a Marie Curie International Reintegration Grant (FP6-039722) and Generalitat Valenciana (GV/2007/065). H.D. acknowledges support from the Generalitat Valenciana (GV06/080) and Ministerio de Educación y Ciencia, Spain (BFU2006-15413-C02-02/BMC). R.C. acknowledges Fondo per gli Investimenti della Ricerca di Base (FIRB) 2003 LIBI-International Laboratory of Bioinformatics (Ministerio dell' Universitá e della Ricerca [MIUR-Italy]) and Network of Excellence (NE) BIOSAPIENS (EU contract number LSHG-CT-2003-503265).

REFERENCES

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J Mol Biol 215:403–410.

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.

Anishetty S, Anishetty R, Pennathur G. 2006. Understanding mutations and protein stability through tripeptides. FEBS Lett 580:2071–2080.

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.

Bairoch A, Apweiler R, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, Martin MJ, Natale DA, O'Donovan C, Redaschi N, Yeh LS. 2005. The Universal Protein Resource (UniProt). Nucleic Acids Res 33(Database issue):D154–D159.

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. 2005. 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. 2005. nsSNPAnalyzer: identifying disease-associated nonsynonymous single nucleotide polymorphisms. Nucleic Acids Res 33(Web Server issue):W480–W482.

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

Capriotti E, Fariselli P, Casadio R. 2005b. I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure. Nucleic Acids Res 33(Web Server issue):W306–W310.

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, Lane CR, 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.

Chan PA, Duraisamy S, Miller PJ, Newell JA, McBride C, Bond JP, Raevaara T, Ollila S, Nystrom M, Grimm AJ, Christodoulou J, Oetting WS, Greenblatt MS. 2007. Interpreting missense variants: comparing computational methods in human disease genes CDKN2A, MLH1, MSH2, MECP2, and tyrosinase (TYR). Hum Mutat 28:683–693.

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

Collins FS, Brooks LD, Chakravarti A. 1998. A DNA polymorphism discovery resource for research on human genetic variation. Genome Res 8:1229–1231.

De Vries EM, Ricke DO, De Vries TN, Hartmann A, Blaszyk H, Liao D, Soussi T, Kovach JS, Sommer SS. 1996. Database of mutations in the p53 and APC tumor suppressor genes designed to facilitate molecular epidemiological analyses. Hum Mutat 7:202–213.

Edgar RC. 2004. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics 5:113.

Hopwood JJ, Bunge S, Morris CP, Wilson PJ, Steglich C, Beck M, Schwinger E, Gal A. 1993. Molecular basis of mucopolysaccharidosis type II: mutations in the iduronate-2-sulphatase gene. Hum Mutat 2:435–442.

Hubbard T, Andrews D, Caccamo M, Cameron G, Chen Y, Clamp M, Clarke L, Coates G, Cox T, Cunningham F, Curwen V, Cutts T, Down T, Durbin R, Fernandez-Suarez XM, Gilbert J, Hammond M, Herrero J, Hotz H, Howe K, Iyer V, Jekosch K, Kahari A, Kasprzyk A, Keefe D, Keenan S, Kokocinsci F, London D, Longden I, McVicker G, Melsopp C, Meidl P, Potter S, Proctor G, Rae M, Rios D, Schuster M, Searle S, Severin J, Slater G, Smedley D, Smith J, Spooner W, Stabenau A, Stalker J, Storey R, Trevanion S, Ureta-Vidal A, Vogel J, White S, Woodwark C, Birney E. 2005. Ensembl 2005. Nucleic Acids Res 33(Database issue):D447–D453.

Ihaka R, Gentelman R. 1996. R: a language for data analysis and graphics. J Comput Graph Stat 5:299–314.

Karchin R, Diekhans M, Kelly L, Thomas DJ, Pieper U, Eswar N, Haussler D, Sali A. 2005a. LS-SNP: large-scale annotation of coding non-synonymous SNPs based on multiple information sources. Bioinformatics 21:2814–2820.

Karchin R, Kelly L, Sali A. 2005b. Improving functional annotation of non-synonymous SNPs with information theory. Pac Symp Biocomput p. 397–408.

Karsten SL, Voskoboeva E, Carlberg BM, Kleijer WJ, Tsnnesen T, Pettersson U, Bondeson ML. 1998. Identification of 9 novel IDS gene mutations in 19 unrelated Hunter syndrome (mucopolysaccharidosis Type II) patients. Mutation in Brief #202. Online. Hum Mutat 12:433.

Kimura M. 1983. The neutral theory of molecular evolution. Cambridge, UK: Cambridge University Press. p. 307–327.

Krawczak M, Ball EV, Fenton I, Stenson PD, Abeysinghe S, Thomas N, Cooper DN. 2000. Human gene mutation database—a biomedical information and research resource. Hum Mutat 15:45–51.

Li W, Godzik A. 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22:1658–1659.

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.

Nielsen LL, Lipari P, Dell J, Gurnani M, Hajian G. 1998. Adenovirus-mediated p53 gene therapy and paclitaxel have synergistic efficacy in models of human head and neck, ovarian, prostate, and breast cancer. Clin Cancer Res 4:835–846.

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

Santibanez Koref MF, Gangeswaran R, Santibanez Koref IP, Shanahan N, Hancock JM. 2003. A phylogenetic approach to assessing the significance of missense mutations in disease genes. Hum Mutat 22:51–58.

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.

Springer MS, Stanhope MJ, Madsen O, de Jong WW. 2004. Molecules consolidate the placental mammal tree. Trends Ecol Evol 19:430–438.

Stenson PD, Ball EV, Mort M, Phillips AD, Shiel JA, Thomas NS, Abeysinghe S, Krawczak M, Cooper DN. 2003. Human Gene Mutation Database (HGMD): 2003 update. Hum Mutat 21:577–581.

Terp BN, Cooper DN, Christensen IT, Jorgensen FS, Bross P, Gregersen N, Krawczak M. 2002. Assessing the relative importance of the biophysical properties of amino acid substitutions associated with human genetic disease. Hum Mutat 20:98–109.

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

Thomas PD, Kejariwal A, Campbell MJ, Mi H, Diemer K, Guo N, Ladunga I, Ulitsky-Lazareva B, Muruganujan A, Rabkin S, Vandergriff JA, Doremieux O. 2003b. PANTHER: a browsable database of gene products organized by biological function, using curated protein family and subfamily classification. Nucleic Acids Res 31:334–341.

Vapnik V. 1995. The nature of statistical learning theory. Berlin: Springer. p 123–167.

Yang Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci 13:555–556.

Yang Z. 2003. Adaptive molecular evolution. New York: Wiley. p 224–254.

Yang Z, Wong WS, Nielsen R. 2005. Bayes empirical Bayes inference of amino acid sites under positive selection. Mol Biol Evol 22:1107–1118.

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