BIOINFORMATICS ORIGINAL PAPER

Vol. 21 no. 12 2005, pages 2814–2820
doi:10.1093/bioinformatics/bti442

Genome analysis

LS-SNP: large-scale annotation of coding non-synonymous SNPs based on multiple information sources

Rachel Karchin[1,2,3,*], Mark Diekhans[4,5], Libusha Kelly[1,2,3], Daryl J. Thomas[4,5], Ursula Pieper[1,2,3], Narayanan Eswar[1,2,3], David Haussler[4,5,6] and Andrej Sali[1,2,3]

[1]Department of Biopharmaceutical Sciences, [2]Department of Pharmaceutical Chemistry and [3]California Institute for Quantitative Biomedical Research, University of California at San Francisco, San Francisco, CA 94143, USA and [4]Center for Biomolecular Science and Engineering, [5]California Institute for Quantitative Biomedical Research, [6]Howard Hughes Medical Institute, University of California, Santa at Cruz, Santa Cruz, CA 95076, USA

Received on February 9, 2005; revised on April 5, 2005; accepted on April 6, 2005
Advance Access publication April 12, 2005

ABSTRACT

Motivation: The NCBI dbSNP database lists over 9 million single nucleotide polymorphisms (SNPs) in the human genome, but currently contains limited annotation information. SNPs that result in amino acid residue changes (nsSNPs) are of critical importance in variation between individuals, including disease and drug sensitivity.

Results: We have developed LS-SNP, a genomic scale software pipeline to annotate nsSNPs. LS-SNP comprehensively maps nsSNPs onto protein sequences, functional pathways and comparative protein structure models, and predicts positions where nsSNPs destabilize proteins, interfere with the formation of domain–domain interfaces, have an effect on protein–ligand binding or severely impact human health. It currently annotates 28 043 validated SNPs that produce amino acid residue substitutions in human proteins from the SwissProt/TrEMBL database. Annotations can be viewed via a web interface either in the context of a genomic region or by selecting sets of SNPs, genes, proteins or pathways. These results are useful for identifying candidate functional SNPs within a gene, haplotype or pathway and in probing molecular mechanisms responsible for functional impacts of nsSNPs.

Availability: http://www.salilab.org/LS-SNP

Contact: rachelk@salilab.org

Supplementary information: http://salilab.org/LS-SNP/supp-info.pdf

INTRODUCTION

With the sequencing of the human genome and recent advances in single nucleotide polymorphism (SNP) detection technologies, numerous coding region polymorphisms that result in amino acid residue changes [i.e. non-synonymous cSNPs (nsSNPs)] have been identified. These SNPs are an important source of interindividual human variation, including disease susceptibility and drug sensitivity. A key challenge in the next few years is to annotate nsSNPs that alter protein function and to link the mechanisms involved to a larger network of intermolecular interactions.

An nsSNP can alter protein function by changing the stability of its native structure and/or its binding properties. Changes in stability may affect a protein's folding rate and increase its susceptibility to proteolysis, resulting in reduced concentration of the native protein. For this reason, several studies have attempted to predict functional impact by mapping nsSNPs onto protein structures and to discover properties of structural context that distinguish disease-associated and neutral nsSNPs (Stitziel *et al*., 2003; Sunyaev *et al*., 2001; Wang and Moult, 2001).

Rule-based approaches for predicting nsSNP functional impact have considered structural context alone (Wang and Moult, 2001) or integrated structural context with evolutionary conservation and statistical information about amino acid hydrophobicity (Sunyaev *et al*., 2001). Rules are commonly benchmarked by comparison of known disease-associated and putatively neutral nsSNPs (or amino acid residue substitutions observed between human proteins and orthologs in other mammals).

One study concluded that most inherited monogenic disease is the direct result of detectable protein defects. Most disease-associated nsSNPs were found to be slightly destabilizing, rather than directly impacting key catalytic or ligand-binding amino acid residues (Wang and Moult, 2001). As described, this method is not suited for large-scale studies. Many of the rules depend on high-resolution atomic structures and required manual structural modeling.

Another study used a rule set to estimate the number of damaging nsSNPs in the average human genome. They concluded that the average human genotype has ∼2000 mildly damaging nsSNPs and provided a list of 158 sample nsSNPs predicted to be damaging (Sunyaev *et al*., 2001).

A third study applied the alpha-shape method (Liang *et al*., 1998) to characterize three distinct structural contexts for nsSNPs: in a pocket or void, in shallow depressed regions or completely buried (Stitziel *et al*., 2003). The majority of disease-associated nsSNPs were found to be in voids and pockets, and infrequently in the interior. When disease-associated nsSNPs located in the protein interior do occur, they are likely to be at highly conserved positions. The method was applied only to proteins with high-resolution experimentally determined structures, presumably because the definition of the geometric features requires atomic level accuracy.

These studies were done on relatively small datasets and yielded a limited number of predicted deleterious nsSNPs. In this study, we attempt to apply all available sources of information to annotate all nsSNPs in the dbSNP database. To scale up the annotation

[*]To whom correspondence should be addressed.

process, we combine rule-based predictions with a machine learning approach, using predictive features that can be automatically extracted from protein sequences, alignments and structures.

The best source of protein structural information is a high resolution, X-ray crystallographic structure, but currently the number of such structures is limited. In August 2004, there were 4404 X-ray structures of human proteins in PDB, compared with 56 275 human protein sequences in the SwissProt/TrEMBL protein database (Boeckmann *et al*., 2003) (including alternatively spliced variants). To enrich available structural data, we use here computationally generated comparative protein structure models (homology-based models) of proteins, built by inferring the structure of a query protein (called the target) from the structure of a putatively homologous protein solved by experimental methods such as X-ray crystallography or nuclear magnetic resonance spectroscopy (NMR) (the template).

We describe a genomic scale, computational pipeline that maps human SNPs in NCBI's dbSNP database (Sherry *et al*., 2001) onto protein sequences in the SwissProt/TrEMBL databases. SwissProt is a manually curated protein sequence database with extensive annotation, and TrEMBL is a computer annotated supplement of SwissProt that contains translations of nucleotide sequence entries not yet integrated into SwissProt. For each sequence, we construct a multiple sequence alignment (MSA) with an iterative database search (Karplus *et al*., 1998). We also build a comparative structure model with an automated procedure that includes fold assignment, target–template alignment, model building and model assessment (Sanchez and Sali, 1998). The resulting collection of sequences, fold assignments, alignments, models and model assessments is used to identify nsSNPs that putatively effect protein function.

The pipeline produced structural annotations for 13 062 dbSNP rsIDs, of which 4907 are dbSNP-validated rsIDs, an increase of several fold with respect to other available online resources, such as ModSNP and PolyPhen. Improved structural coverage of SNPs and integration with databases of protein–protein interactions, MODBASE (Pieper *et al*., 2004) and PIBASE (Davis and Sali, 2005) makes it possible to explore how SNPs in a gene, haplotype or pathway might interact on a molecular level.

METHODS

We implemented an automated computational pipeline consisting of three modules (Fig. 1). In the first module, we extract the genomic locations of human SNPs from dbSNP and map these SNPs onto human protein sequences in SwissProt/TrEMBL to identify the SNPs that result in an amino acid residue substitution. In the second module, for each SwissProt/TrEMBL protein sequence (target), we identify related known protein structures (templates), align each target to the template, and build as well as assess corresponding comparative structure models. In the third module, we use amino acid residue substitutions, MSAs and comparative models to identify nsSNPs that generally impact human health and specifically nsSNPs that interfere with the formation of domain–domain interfaces or have an effect on protein–ligand binding. These three modules are described next.

Module 1: SNP-to-protein mapping

Although dbSNP provides mappings of SNPs onto protein coding regions, these mappings are restricted to proteins in reference sequences that have been curated in NCBI's LocusLink (Maglott *et al*., 2000). To enrich the number of proteins in our dataset, we did our own mappings based on the KnownGenes mRNA-to-genome alignments available from the UCSC Genome Browser (July 2003 hg16, NCBI Build 34) (Hsu *et al*., 2005; Kent *et al*., 2002).

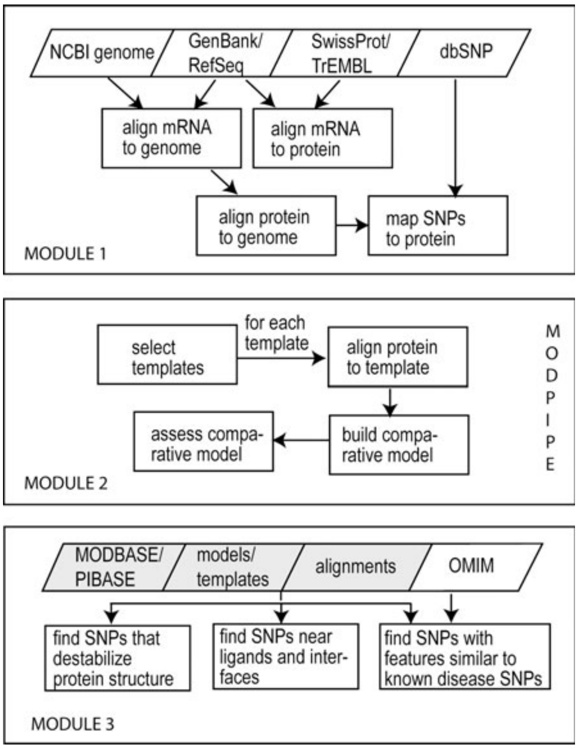

Fig. 1. The LS-SNP computational pipeline maps SNPs onto protein sequences, builds comparative structure models and annotates SNPs using features of protein structure, sequence and evolution.

The genomic positions of all human SNPs were obtained from dbSNP build 120. From a total of 9.1 million SNP reference clusters (rsIDs), we eliminated 580 000 indels, 32 000 segmental mutations and 1.9 million SNPs that map to more than two genomic positions.

To map SNPs to the amino acid residues of a protein, each amino acid residue was aligned to three genomic base positions using the protein's representative mRNA. First, the protein was aligned to its mRNA with TBLASTN (Altschul *et al*., 1990). Partially aligned or unaligned amino acid residues were ignored. The protein-to-mRNA alignment was then mapped to the genome using mRNA-to-genomic DNA alignments produced by BLAT (Kent, 2002). This approach compensates for cases where the representative mRNA does not exactly code for the protein, due to mRNA editing and reverse transcriptase or mRNA sequencing errors (Furey *et al*., 2004). Each mRNA/protein pair was examined for SNPs that cause amino acid residue substitutions.

The primary output of the SNP-to-protein mapping module is a list of protein sequences from SwissProt/TrEMBL and the positions of all amino acid residue substitutions produced by the SNPs found in dbSNP.

Module 2: from sequence to structure

The SwissProt/TrEMBL protein sequences were input into MODPIPE, an automated system for comparative protein structure modeling (Eswar *et al*., 2003; Sanchez and Sali, 1998). MODPIPE generally calculates comparative models for a protein sequence using a number of different template structures and sequence–structure alignments. Sequence–structure matches are identified by aligning the PSI-BLAST profile of each sequence (built with 10 iterations and E-value cutoff 0.0001) against a library of candidate template sequences extracted from PDB and by scanning the sequence against a database of template profiles with IMPALA (Schaffer *et al*., 1999). Each significant alignment (E-value cutoff 0.0001) that covers distinct regions of the target sequence is chosen for modeling. Models are calculated for each of the sequence–structure matches using the default 'model' routine of MODELLER (Sali and Blundell, 1993). A statistical scoring function is used to assess each model (Melo *et al*., 2002).

For each amino acid residue substitution found in a protein sequence, we performed *in silico* mutation of its comparative model by applying the 'mutate_model' routine in MODELLER, followed by a combination of conjugate gradient minimization and molecular dynamics with simulated annealing of the mutated residue and its neighbors (E. Feyfant, personal communication).

The output of the sequence-to-structure module is a collection of fold assignments, alignments of target sequences and template structures, comparative structure models for SwissProt/TrEMBL sequences and mutated sequences, and model assessments.

Module 3: annotations

We use the output of the first two modules to help compute a variety of annotations for human nsSNPs. The nsSNPs are annotated with respect to genomic sequence, protein sequence, protein structure and function. Our webserver allows a user to view the annotations from different viewpoints: a single nsSNP, all nsSNPs in a protein sequence or gene of interest, all nsSNPs within a genomic region of interest and all nsSNPs in a functional pathway of interest. Finally, we combine a rule-based approach to identify putatively destabilizing nsSNPs and a supervised machine learning approach to identify nsSNPs likely to have an impact on human health. The webserver annotations are linked to the KEGG (Ogata *et al*., 1999), SwissProt, SCOP (Murzin *et al*., 1995), PDBSUM (Laskowski *et al*., 1997) and PIBASE databases, and cross-linked with MODBASE and UCSC Genome Browser.

Rule-based annotations To identify changes that may destabilize protein structure, we applied four structural rules that are based on the preferences of each amino acid residue type to be in any of the secondary structure and solvent accessibility states (Sunyaev *et al*., 2001).

We used the DSSP program (Kabsch and Sander, 1983) to compute secondary structure state and solvent accessible surface area at each position where an nsSNP occurs in our comparative models. Relative solvent accessibility (RSA) was calculated by normalizing with respect to maximum solvent accessibility for each residue type (Rost and Sander, 1994). The accessible surface propensity tables were downloaded from the PolyPhen web site.

Destabilization is predicted when any one of the following conditions is satisfied: (1) RSA is $<25\%$ and difference in accessible surface propensities is >0.75; (2) RSA is $>50\%$ and difference in accessible surface propensities is >2; (3) RSA is $<25\%$ and formal charge change (histidine is assigned a $+1$ charge); (4) the variant involves a proline in a helix (Sunyaev *et al*., 2001).

Interference with domain–domain interface formation or protein–ligand binding is predicted when any of the four conditions listed above occur at a putative domain–domain interface or ligand binding site. To find such nsSNPs, we identified template residues at domain–domain interfaces and in proximity to small molecule ligands using PIBASE and the LIGBASE table (Stuart *et al*., 2002) of MODBASE, respectively. A template residue is considered to be at an interface if it is within 6 Å of an atom in an adjacent domain. It is considered to be ligand binding if it is within 5 Å of a HETATM (i.e. an atom not covalently bonded to the protein, not in one of the standard 20 residue types, nor in a water molecule) in the PDB structure. For each nsSNP position within a structurally modeled region of protein sequence, we used the sequence-to-template alignments to specify the equivalent residue position in the template structure. An nsSNP position is putatively ligand binding if it aligns to a ligand-binding template residue. Similarly, it is predicted to be at a domain–domain interface if it aligns to an interface template residue.

Supervised learning To discriminate between disease-associated nsSNPs on one hand and neutral or positive nsSNPs on the other hand, we applied a supervised machine learning approach that combines information from multiple sources: amino acid residue side chain properties, comparative structure models of the SwissProt/TrEMBL sequences and mutated sequences and evolutionary properties extracted from MSAs (Ng and Henikoff, 2002).

Table 1. List of 13 features used to train an SVM to discriminate between monogenic disease-associated amino acid residue substitutions from OMIM and neutral/positive non-synonymous SNPs from dbSNP

Category	Feature
Protein structure modeling	Fractional solvent accessibility (wild-type)
	Fractional solvent accessibility (mutant)
	Solvent accessibility (wild-type)
	Solvent accessibility (mutant)
	In silico mutation violated spatial restraints (z-score)
	In silico mutation violated spatial restraints (molecular pdf)
	Buried charge
Evolution	Hidden Markov model PHC score
	Hidden Markov model relative entropy
Amino acid residues	Grantham values
	Change in residue volume
	Change in residue hydrophobicity
	Change in residue formal charge

To compute evolutionary properties based on amino acid residue conservation and substitution likelihoods, we constructed an MSA for each SwissProt/TrEMBL protein sequence, via iterative search of NCBI's nr database using the SAM-T2K algorithm (Karplus *et al*., 1998).

A benchmark set of 1457 disease-associated and 2504 putatively neutral nsSNPs was assembled as follows. We found each amino acid residue substitution in a protein sequence that is annotated as VARIANT in the SwissProt database. VARIANT substitutions with corresponding entries in OMIM (McKusick, 2000) were assigned to the disease-associated class. We assigned all validated dbSNP nsSNPs that were not annotated as destabilizing (according to rules 1–4 above) to the putatively neutral class.

For each nsSNP, we computed 13 properties (features) using a variety of programs, including MODELLER, MODPIPE (Sanchez and Sali, 1998), DSSP (Kabsch and Sander, 1983), and SAM (Krogh *et al*., 1994) (Table 1) as described previously (Karchin *et al*., 2005). We add here two new features that measure strain when a mutant side chain is introduced into the native sequence; the strain is quantified by the number of violated spatial restraints used in the construction of the mutant model.

We trained a support vector machine (SVM) (Vapnik, 1995) to discriminate between the two classes of nsSNPs. SVM kernel and parameter selection was done using a 3-fold cross-validation protocol. The benchmark dataset was randomly partitioned into three subsets. The SVM was trained on subsets 1 and 2, tested on subset 3, trained on 1 and 3, tested on 2, etc. We report an overall prediction accuracy (percentage of nsSNPs correctly predicted), false positive rate (percentage of misclassified disease-causing variants), and false negative rate (percentage of misclassified neutral or positive nsSNPs) by averaging results on the three test subsets. Error bars were obtained by repeating the cross-validation experiments 10 times.

The SVM classifies each example with a discriminant score. In our implementation, negative scores predict disease association while positive scores predict a neutral or positive nsSNP. The absolute value of the score provides a confidence measure for the prediction. All feature selection and SVM learning was done with inhouse software tools coded in PERL and Java.

Owing to the poor performance of a linear SVM on the benchmark (data not shown) we chose to use non-linear kernels. We expect that some examples in our benchmark are mislabeled and use a soft margin, which tolerates noisy labels (Cortes and Vapnik, 1995). The best performance was obtained with a radial basis kernel function and 1-norm soft margin with parameter $C = 1$. We get the highest accuracy and smallest error bars by excluding low confidence predictions where the absolute value of the discriminant score is <0.2.

For the benchmark 3-fold cross-validation test, the SVM achieves prediction accuracy of 80.5% (±0.3%), false positive rate of 19.7% (±0.2%) and false negative rate of 18.7% (±0.8%). The number of rejected, low-confidence predictions is 122 (±3), corresponding to only 3% of the examples in the benchmark.

Using these parameters, we trained an SVM on the entire benchmark set without partitioning. We then applied the optimized SVM to classify the validated nsSNPs occurring in structurally modeled regions of protein sequence.

RESULTS

The SNP-to-protein mapping identified 70 153 coding region SNPs in dbSNP that produce amino acid residue substitutions in 24 944 protein sequences (Supplementary Table 1). Of these SNPs, 28 043 in 14 551 proteins are assigned 'validated' status by dbSNP. The quality of the protein-to-genome alignments is supported by an average 98.8% sequence identity between amino acid residues and translated codons.

The dbSNP database uses an in-house SNP-to-protein mapping that identifies 54 048 coding, non-synonymous human SNPs, of which 21 255 are validated (dbSNP build 120). Using the genomic location of these SNPs (from dbSNP) and our own protein-to-genome alignments, we identify 28 043 validated nsSNPs, an increase of 31.9% compared with the assignments by NCBI. If non-validated SNPs are included, we identify 70 147 SNPs, an increase of 30.8% relative to NCBI. The genomic sequence details (SNP address, chromosome, codon position, strand and nucleotide substitution) are accessible via our webserver by querying with information type set to 'Genomic sequence'. Protein sequence details (SwissProt/TrEMBL AC identifier, amino acid residue changes and residue position) for each nsSNP are available by setting information type to 'Protein sequence'.

We were able to build comparative models for domains in 13 391 proteins, 53% of the proteins in the genome-to-protein alignments (Supplementary Table 2). For the remaining 47%, a protein of known structure with sufficient similarity was not available in the PDB. As described in the Methods section, a protein may have several models, each covering a distinct region of its sequence. For our purposes, we are interested primarily in models that cover an nsSNP position. A covered position is one in which we can identify an equivalent residue in the structure of the template protein used to build the model. The equivalent residue aligns to the nsSNP position in the target–template alignment.

If we consider both validated and non-validated SNPs, our set contains 40777 nsSNP-covering models of 8931 proteins, covering validated 13062 nsSNPs. We have 8725 models of 4593 proteins covering 4907 validated nsSNPs. Sixty-seven of these nsSNPs appear in more than one protein sequence, primarily due to alternative splicing. The average sequence identity in the target–template alignments used to build the models is 28.5% (27.8% if only validated nsSNPs are considered). All alignments have statistically significant Blast E-values ≤ 0.0001.

Structural annotations extracted from the comparative models and target–template alignments are accessible via our webserver by querying with information type set to 'Protein structure'. The server lists DSSP secondary structure, relative solvent accessibility, template PDB code, the equivalent template residue that aligns to the SNP in target–template alignments, model assessment score and target–template sequence identity. For nsSNPs in the proximity of a ligand, we include the ligand name. For nsSNPs near a domain interface, we include SCOP domain identifiers of the interacting domains [Examples in Supplemenatry Table 3(a)].

These structural annotations allow us to identify 1886 putatively destabilizing nsSNPs with the structural rules described in the Methods section [Examples in Supplementary Table 3(b)]. A total of 1317 putative monogenic disease-associated nsSNPs are predicted by an SVM trained on properties extracted from the comparative models, SAM-T2K MSAs and properties of amino acid residue substitutions [Examples in Supplementary Table 3(c)]. The functional annotations are accessible from the webserver by querying with information type set to 'Functional'.

The LS-SNP webserver is a front-end to a relational (mySQL) database. All information is precalculated by an automated build procedure. We plan to update once or twice a year to synchronize with significant releases of dbSNP. Batch queries via an external program can be implemented with URL strings containing a genomic range or an arbitrary number of genes, pathways or SNPs (instructions in online Help).

Case studies

To illustrate the utility of LS-SNP, we describe next three proteins containing nsSNPs that we have annotated as being destabilizing and/or disease-associated (Supplementary Figure 1). We have found confirmation for two of the annotations in the clinical literature.

Dihydropyrimidine dehydrogenase Human dihydropyrimidine dehydrogenase (DPD) contains an SNP (dbSNP ID rs1801266 Arg235Trp) that causes a buried charge change and whose equivalent residue is close to a ligand. This enzyme catalyzes NADPH-dependent reduction of uracil and thymine, the first and rate-limiting step in pyrimidine degradation. It also degrades the cancer chemotherapeutic 5-fluoro-uracil. The template for the comparative model is the PDB structure 1gte, a DPD from pig, complexed with FAD in place of NADPH. Overall target–template sequence identity is 93%, with 97.26% sequence identity in the region around the ligand. Model assessment score is 1.0. The Arg235Trp mutation has been associated with DPD deficiency in the clinical literature (Vreken *et al*., 1997), although a published study of DPD structure–disease correlations overlooked it (van Kuilenburg *et al*., 2002). This case is an example of a clinically documented SNP that is not validated in dbSNP. Because LS-SNP is comprehensive, it can help users identify deleterious SNPs that have been missed in small-scale studies. We functionally annotate the Arg235Trp nsSNP and identify its biological/medical relevance by integrating information from comparative modeling, annotation of the target protein (from SwissProt) and the template protein (from LIGBASE and PDBSUM), and application of structural rules.

Glutathione S-transferase Glutathione *S*-transferases (GSTs) play a key role in cellular detoxification by conjugating the tripeptide glutathione (GSH) to a variety of xenobiotics. They are associated with cellular resistance to anticancer drugs, insecticides, herbicides and antibiotics (Rossjohn *et al*., 1998). We identified a SNP (dbSNP ID rs2234953) in GST theta 1 (Glu172Lys) that lies at a domain interface position. It produces a buried charge change and an unfavorable change in accessible surface potential at a buried position. The model is based on human GST theta 2 (PDB code 1ljr). Target–template sequence identity is 51% (67% if only interface residues are considered). Model assessment score is 1.0. In GST theta 2, Glu172 and Arg107 are in close proximity and Arg107 interacts with the thiol sulfur of the GSH ligand (Rossjohn *et al*., 1998). These sidechains are conserved in GST theta 1, and in GST-theta from mouse, rat and chicken (Flanagan *et al*., 1998), supporting the idea that the Arg107Glu substitution may be deleterious. The functional annotation is based on a variety of information sources brought together through LS-SNP that might not be convincing individually (comparative structure model with 51% target–template sequence identity, domain interface location from PIBASE, structural rules, conserved sidechains). Taken as a whole, the information sources reinforce each other and suggest a molecular mechanism underlying putative deleterious effects.

Cytochrome P450 2A6 Cytochrome P450 2A6 (also named 2A3) is an enzyme involved in oxidation of nicotine and coumarin in human liver microsomes and has been associated with nicotine addiction and lung cancer susceptibility (Pianezza *et al*., 1998; Yamano *et al*., 1990). We found a SNP (dbSNP rsID 1801272, Leu160His) that produces a buried charge change. The model is based on rabbit cytochrome P450 2C5 (PDB code 1nr6). Overall target–template sequence identity is 51%, and the Leu is at a conserved position in the alignment. Model assessment score is 1.0. The clinical literature confirms that this SNP produces an unstable and catalytically inactive enzyme (Yamano *et al*., 1990). The cytochrome P450 2C5 and 2A6 have different functional specificity, but because the model is based on the correct fold (overall topology) and a good alignment, it provides useful information about the SNP.

DISCUSSION

Data quality

Our source of SNP data is NCBI's dbSNP, a public database with an open submission policy. The dbSNP definition of SNP is any single nucleotide polymorphism, regardless of its allelic frequency. The data are the product of many genotyping methods, which have varying reliabilities. For example, SNPs identified through cDNA-mediated sequencing of mRNA with reverse transcriptase (RT–PCR) are more likely to be false positives than those identified through direct genomic sequencing, owing to the high error rate of reverse transcriptase (Furey *et al*., 2004; Gerhard *et al*., 2004). As a quality control, dbSNP distinguishes between validated and unvalidated entries. A SNP can be validated by multiple, independent submissions; frequency/genotype data; submitter confirmation; alleles observed in at least two chromosomes or submission by the HapMap project. Approximately 50% of the SNPs are validated (Sherry *et al*., 2001).

We considered including only the SNPs that are assigned 'validation' status by dbSNP, or more conservatively, those with frequency/genotype data. However, there are many examples of non-validated SNPs described in the clinical literature (Frayling *et al*., 1998; Lamlum *et al*., 2000; Nijbroek *et al*., 1995; Ory *et al*., 1989). Therefore, to maximize our coverage of putative nsSNPs, we have annotated both validated and non-validated SNPs. Our website provides an option to filter non-validated entries from user's query results.

We frequently have multiple annotations for a single SNP. In our DNA-to-protein mappings, one SNP may map to several SwissProt/TrEMBL proteins because of alternative splicing events or because the protein sequence appeared with more than one accession number in SwissProt/TrEMBL at the time of our build. A SNP may also be covered by several models. Often, we have several models for a protein based on different PDB template structures. A SNP may also be found in multiple models based on the same template, if different overlapping segments of the target produce statistically significant PSI-BLAST and/or IMPALA alignments with the same template.

Ligand annotations

Although we annotate 1439 nsSNPs (and 443 validated nsSNPs) as being close to a ligand, some of these ligands may not be biologically interesting. For example, crystallization reagents such as glycerol, ethylene glycol, ethanol and inositol may appear as non-covalently bonded HETATMs in PDB files. In most cases, the appearance of a SNP in the proximity of glycerol will have no functional effects, but in glycerol kinase (an enzyme in which glycerol is the substrate), the SNP could be important. Since our automated methods cannot yet evaluate the local environment of each SNP-ligand contact, we report all such contacts and let the user decide which are of interest.

Structural annotations

The protein structure models used in this work are based on target–template alignments with average sequence identity of 28.5%. We have not applied any cutoff based on target-template sequence identity. Instead, we make available all of our data and provide users with two measures of confidence to consider when looking at our structure-based annotations: target–template sequence identity and model assessment score. The assessment score ranges from 0 for models that tend to have an incorrect fold to 1.0 for models that tend to be comparable at least to low-resolution X-ray structures. Comparison of models with their corresponding experimental structures indicates that models with scores >0.7 generally have the correct fold, with $>35\%$ of the backbone atoms superposable within 3.5 Å. Reliable models (score ≥ 0.7) based on alignments with $>40\%$ sequence identity have a median overlap of $>90\%$ with the corresponding experimental structure. In the 30–40% sequence identity range, the overlap is usually between 75 and 90% and below 30% it drops to 50–75%, or even less in the worst cases (Eswar *et al*., 2003). It is also worth considering that even when overall target–template sequence identity is low, sequence identities in regions of functional importance, such as binding sites, may be higher.

Because our structural annotations are based on comparative models, rather than high-resolution crystallographic structures, we were careful to use properties that depend primarily on correct fold assignment and a good target–template alignment (in the region of the nsSNP), as opposed to atomic-level structural details such as loss of salt bridges, hydrogen or disulfide bonds.

In contrast to our approach, some studies have used stringent quality measures, with the trade-off of fewer structurally annotated SNPs. The initial Polyphen survey considered only curated nsSNPs from the HGVBase database and 50% sequence identity cutoff in their alignments with proteins of known structure. Currently, PolyPhen has structural annotations for 3167 dbSNP rsIDs (Ramensky *et al*., 2002). SwissProt's ModSNP database uses a 70% sequence identity cutoff and requires that the wild-type amino acid residue at an nsSNP position exactly match the aligned template amino acid residue. ModSNP has structural annotations for 4109 missense mutations (Yip *et al*., 2004). LS-SNP has structural annotations for 13 062 dbSNP rsIDs, of which 4907 are validated.

For our prediction of disease-associated nsSNPs, we use structural features extracted from *in silico* mutation by the 'mutate_model' routine in MODELLER. Single point mutations can produce a wide range of effects on protein structure, from radical changes in backbone geometry to small sidechain alterations (reviewed in Eigenbrot and Kossiakoff, 1992). Unfortunately, our current methods are not able to accurately model these effects. Nevertheless, useful information for predicting the functional effects of point mutations can still be extracted from an *in silico* mutation procedure, according to a mutual information analysis (Karchin *et al*., 2005). Specifically, this information includes the solvent accessible surface areas of the wild-type and mutant amino acid residues as well as the number of spatial restraint violations after refinement of the mutant.

Large-scale computational prediction of functional SNPs enables a direct approach for identifying SNPs to be genotyped and tested in candidate–gene association studies, by narrowing down the number of SNPs to be looked at in a region. It can also be used to enhance indirect methods in which common tag SNPs are initially used to locate genomic regions containing causal variants. Detection of causal variants is important, given that the promise of personalized medicine is not only to identify high-risk individuals for particular diseases but to develop targeted therapeutics. However, a limitation of our current approach is that deleterious effects on protein stability and binding are necessary but not sufficient conditions for human disease.

As the number of known SNPs increases, so does the value added by automated, large-scale annotation methods. We are currently working on improved template selection, alignment, model building and model assessment methods. We are also developing a formalism for selecting combinations of features that are most predictive of nsSNP functional effects. Of particular interest to us are features that describe the role of an nsSNP within a larger system of interacting molecules. The next version of LS-SNP will include variants from the HapMap project which became available in dbSNP build 123. We expect that the quality and quantity of our annotations will improve with the growth of the dbSNP, SwissProt and PDB databases, and with advances in our comparative modeling and functional prediction methods.

ACKNOWLEDGEMENTS

The authors would like to thank Jim Kent and Fan Hsu for developing Known Genes, Fred Davis for PIBASE, and Ash Stuart and Andrea Rossi for LIGBASE. The project has been supported by NIH F32 GM-072403-01, NIH U01 GM-61390-04, NHGRI P41 HG02371, NCI 22XS013A, HHMI, Sun Academic Equipment Grant EDUD-7824-020257-US, an IBM SUR grant, an Intel computer hardware gift, and the Sandler Family Supporting Foundation.

REFERENCES

Altschul,S.F. *et al*. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.

Boeckmann,B. *et al*. (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.*, **31**, 365–370.

Cortes,C. and Vapnik,V. (1995) Support vector networks. *Machine Learning*, **20**, 273–297.

Davis,F.P. and Sali,A. (2005) PIBASE: a comprehensive database of structurally defined protein domain interfaces. *Bioinformatics*, Jan 18 [Epub ahead of print].

Eigenbrot,C. and Kossiakoff,A.A. (1992) Structural consequences of mutation. *Curr. Opin. Biotechnol.*, **3**, 333–337.

Eswar,N. *et al*. (2003) Tools for comparative protein structure modeling and analysis. *Nucleic Acids Res.*, **31**, 3375–3380.

Flanagan,J.U. *et al*. (1998) A homology model for the human theta-class glutathione transferase T1-1. *Proteins*, **33**, 444–454.

Frayling,I.M. *et al*. (1998) The APC variants I1307K and E1317Q are associated with colorectal tumors, but not always with a family history. *Proc. Natl Acad. Sci. USA*, **95**, 10722–10727.

Furey,T.S. *et al*. (2004) Analysis of human mRNAs with the reference genome sequence reveals potential errors, polymorphisms, and RNA editing. *Genome Res.*, **14**, 2034–2040.

Gerhard,D.S. *et al*. (2004) The status, quality, and expansion of the NIH full-length cDNA project: the mammalian gene collection (MGC). *Genome Res.*, **14**, 2121–2127.

Hsu,F. *et al*. (2005) The UCSC proteome browser. *Nucleic Acids Res.*, **33** (Database Issue), D454–D458.

Kabsch,W. and Sander,C. (1983) Dictionary of protein secondary structure. *Biopolymers*, **22**, 2577–2637.

Karchin,R. *et al*. (2005) Improving functional annotation of non-synonomous SNPs with information theory. *Pac. Symp. Biocomput*, World Scientific, pp. 397–408.

Karplus,K. *et al*. (1998) Hidden Markov models for detecting remote protein homologies. *Bioinformatics*, **14**, 846–856.

Kent,W.J. (2002) BLAT—the BLAST-like alignment tool. *Genome Res.*, **12**, 656–664.

Kent,W.J. *et al*. (2002) The human genome browser at UCSC. *Genome Res.*, **12**, 996–1006.

Krogh,A. *et al*. (1994) Hidden Markov models in computational biology: applications to protein modeling. *J. Mol. Biol.*, **235**, 1501–1531.

Lamlum,H. *et al*. (2000) Germline APC variants in patients with multiple colorectal adenomas, with evidence for the particular importance of E1317Q. *Hum. Mol. Genet.*, **9**, 2215–2221.

Laskowski,R.A. *et al*. (1997) PDBsum: a web-based database of summaries and analyses of all PDB structures. *Trends Biochem. Sci.*, **22**, 488–490.

Liang,J. *et al*. (1998) Analytical shape computation of macromolecules. Proteins, **33**, 1–17.

Maglott,D.R. *et al*. (2000) NCBI's LocusLink and RefSeq. *Nucleic Acids Res.*, **28**, 126–128.

McKusick,V. (2000) Online Mendelian Inheritance in Man, OMIM, Nathans Institute for Genetic Medicine.

Melo,F. *et al*. (2002) Statistical potentials for fold assessment. *Protein Sci.*, **11**, 430–448.

Murzin,A.G. *et al*. (1995) SCOP: a structural classification of proteins database. *J. Mol. Biol.*, **247**, 536–540.

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

Nijbroek,G *et al*. (1995) Fifteen novel FBN1 mutations causing Marfan syndrome. *Am. J. Hum. Genet.*, **57**, 8–21.

Ogata,H. *et al*. (1999) KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, **27**, 29–34.

Ory,P.A. *et al*. (1989) Sequences of complementary DNAs that encode the NA1 and NA2 forms of Fc receptor III on human neutrophils. *J. Clin. Invest.*, **84**, 1688–1691.

Pianezza,M.L. *et al*. (1998) Nicotine metabolism defect reduces smoking. *Nature*, **393**, 750.

Pieper,U. *et al*. (2004) MODBASE, a database of annotated comparative protein structure models, and associated resources. *Nucleic Acids Res.*, **32**, D217–D222.

Ramensky,V. *et al*. (2002) Human non-synonymous SNPs: server and survey. *Nucleic Acids Res.*, **30**, 3894–3900.

Rossjohn,J. *et al*. (1998) Human theta class glutathione transferase: the crystal structure reveals a sulfate-binding pocket within a buried active site. *Structure*, **6**, 309–322.

Rost,B. and Sander,C. (1994) Conservation and prediction of solvent accessibility in protein families. *Proteins*, **20**, 216-226.

Sali,A. and Blundell,T.L. (1993) Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.*, **234**, 779–815.

Sanchez,R. and Sali,A. (1998) Large-scale protein structure modeling of the *S.cerevisiae* genome. *Proc. Natl. Acad. Sci. USA*, **95**, 13597–13602.

Schaffer,A.A. *et al*. (1999) IMPALA: matching a protein sequence against a collection of PSI-BLAST-constructed position-specific score matrices. *Bioinformatics*, **15**, 1000–1011.

Sherry,S.T. *et al*. (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.*, **29**, 308–311.

Stitziel,N.O. *et al*. (2003) Structural location of disease-associated SNPs. *J. Mol. Biol.*, **327**, 1021–1030.

Stuart,A.C. *et al*. (2002) LigBase: a database of families of aligned ligand binding sites in known protein sequences and structures. *Bioinformatics*, **18**, 200–201.

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

van Kuilenburg,A.B. *et al*. (2002) Novel disease-causing mutations in the dihydropyrimidine dehydrogenase gene interpreted by analysis of the three-dimensional protein structure. *Biochem. J.*, **364**, 157–163.

Vapnik,V. (1995) *The Nature of Statistical Learning Theory*. New York: Springer-Verlag.

Vreken,P. *et al*. (1997) Dihydropyrimidine dehydrogenase (DPD) deficiency: identification and expression of missense mutations C29R, R886H and R235W. *Hum. Genet.*, **101**, 333–338.

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

Yamano,S. *et al*. (1990) The CYP2A3 gene product catalyzes coumarin 7-hydroxylation in human liver microsomes. *Biochemistry*, **29**, 1322–1329.

Yip,Y.L. *et al*. (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.