ABSTRACT

Title of Document: COMPUTATIONAL MODELING OF THE RELATIONSHIP BETWEEN SNPS AND DISEASE

Peng Yue, Doctor of Philosophy, 2005

Directed By: Professor, John Moult, CARB/UMBI

We have developed two models, the stability model and the profile model, to identify non-synonymous single base changes (the most common cause of monogenic disease) that have deleterious effects on protein function in vivo. The stability model analyzes the effect of the resulting amino acid change on protein stability by utilizing structural information such as reduction in hydrophobic area and loss of electrostatic interactions. The profile model makes use of the conservation and type of residues observed at a base change position within a protein family. In each model, a machine learning technique, the support vector machine (SVM) was trained on a set of mutations causative of disease, and a control set of non-disease causing mutations. In jack-knifed testing, the stability model identifies 74% of disease mutations, with a false positive rate of 15%; the profile model identifies 80% of disease mutations, with a false positive rate of 10%. Evaluation of a set of in vitro mutagenesis data with the stability model established that the majority of disease mutations affect protein stability by 1 to 3 Kcal/mol. The stability model's effective distinction between disease and non-disease variants strongly supports the hypothesis that loss of protein stability is a major factor contributing to monogenic disease.

Both models are used to identify deleterious SNPs in the human population. After carefully controlling of errors, we find that approximately one-fourth of the known non-synonymous SNPs are deleterious, thus providing a set of possible SNPs contributing to human complex disease traits.

A web resource has been developed to provide information on disease/gene relationships at the molecular level. The resource has three primary modules. The first module is used to publish the deleterious SNPs identified by the two above-mentioned models. The second module identifies the candidate genes for a specific disease, and the third module provides information about the relationships between the sets of candidate genes. Disease/candidate gene relationships and gene-gene relationships are derived from the literature using a simple but effective text profiling method.

COMPUTATIONAL MODELING OF THE RELATIONSHIP BETWEEN SNPS AND DISEASE

By

Peng Yue

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2005

Advisory Committee:
Professor John Moult, Chair
Professor Michael Gilson
Professor Richard Payne
Associate Professor Stephen Mount
Associate Professor Sarah Tishkoff
Associate Professor Victor Muñoz, Dean’s representative

Acknowledgements

I want to thank Dr. John Moult, my mentor and supervisor, from whom I have learned so much over the years. Without his great insight, knowledge, patience and kindness, this dissertation could not have been written. I am also grateful to all my other committee members, Dr. Michael Gilson, Dr. Richard Payne, Dr. Sarah Tishkoff, and Dr. Stephen Mount. Their helpful advice and guidance have certainly played an essential role in improvement of this work. I would also like to thank Dr. Victor Muñoz for being Dean's representative.

I am thankful to Ms. Zhaolong Li, for her contribution in the stability model. My special thanks go to Mr. Eugene Melamud, with whom I have worked closely in the past six years and developed such deep friendship. His unselfish help have contributed tremendously to the success of this project.

Last but not least, my deepest appreciation goes to my family, especially to my wife and my mother. Without their loving support, this long journey toward completion of my study and dissertation is doomed to be a lonely and more arduous one.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

SNP in Populations and Human Disease

A single nucleotide polymorphism (SNP) is a single nucleotide variation in DNA. In the human population, SNPs are the most abundant genetic variations. It is estimated that human population has approximately 10 million common SNPs with minor allele frequency (MAF) of more than 1% (Kruglyak and Nickerson 2001). 50,000 – 100,000 of these common SNPs are non-synonymous SNPs (i.e., resulting in an amino acid change) (Botstein and Risch 2003; Cargill et al. 1999; Halushka et al. 1999). In the current central SNP repository of dbSNP (Sherry et al. 2001), over 10 million SNPs have been deposited, of which five million have been validated.

In theory, a SNP may affect gene function through a number of mechanisms including changes in transcription, RNA processing, protein translation, folding of the polypeptide chain, stability of the folded state, post-translational modification, interactions with binding partners, and alterations to catalysis. An analysis of the Human Gene Mutation Database (HGMD) (Stenson et al. 2003) has shown that the vast majority of known monogenic disease cases act through changes to the coding sequence, with mis-sense mutations (a single base change resulting in change of a single amino acid) accounting for greater than 60% of all monogenic disease mutations. Mis-sense SNPs may also be the major mutation type underlying human complex diseases. The following reasons support this view: first, this type of mutation is prevalent in monogenic disease, and second, in the human genome, coding regions are larger than other functional regions, such as those for transcription regulation and splicing regulation. However, it has been suggested that complex diseases may be predominantly affected by SNPs within gene regulation regions (Hirschhorn and Daly 2005).

Two Types of Human Disease

Over the past 20 years, more than 1000 disease genes have been identified by genetic mapping, especially linkage analysis. Most of these are genes for rare monogenic (one gene/one trait) disease which follow a simple Mendelian inheritance pattern. Common human disease, such as hypertension, diabetes, Alzheimer, stroke, and heart disease, on the other hand, follow a more complicated inheritance pattern. As a consequence, common diseases prove to be much harder to analyze (Botstein and Risch 2003; Carlson et al. 2004; Emahazion et al. 2001). The difficulty in analyzing common diseases may be caused by incomplete penetrance (a person carrying a predisposing allele may not exhibit the disease phenotype), genetic heterogeneity (mutations on one of several genes may result in identical phenotypes), and polygenic inheritance (a trait is controlled by multiple gene interactions so that each individual predisposing allele has a low risk factor and shows weak correlation with the disease trait). In addition, environmental factors may also play an important role in shaping disease phenotypes. Many phenotypic traits, such as behavioral characteristics and different drug response are also believed to follow such a complex inheritance pattern and are thus generally called complex traits.

The Allelic Structure of Human Disease

Owning to the extensive data on human monogenic diseases, their allelic structure is relatively clear. Monogenic diseases are usually very rare in the human population and the frequency of disease alleles is also very low, usually << 1%. The level of allelic heterogeneity is very high in monogenic diseases. According to the human gene mutation database (HGMD) (Stenson et al. 2003), there are on average over 10 disease alleles per disease gene.

However, little is known about the allelic structure of human complex diseases, since very few complex disease genes have been unambiguously identified. One popular view is the common disease/common variant (CD/CV) model (Reich and Lander 2001), which assumes that human common diseases are caused by one or a few predominating alleles at a small number of loci. Those alleles are generally old and common across different geographical populations. Known complex disease alleles seem to support this hypothesis, such as the APOE ε4 allele in Alzhemer's disease (Corder et al. 1993) and the PPARγ P12A in type II diabetes (Altshuler et al. 2000). However, such a small sample is not sufficient to exclude other possible models. For example, complex diseases may be affected by rare susceptibility alleles at a large number of loci. The CD/CV model represents the best case scenario that we could possibly have for identifying disease genes. On this basis, association studies are proposed as a feasible way to identify disease genes, as discussed in detail later.

Genetic Mapping of Human Disease

Linkage Analysis

Linkage analysis is performed on family pedigrees. The extent to which a genetic marker and a disease trait are co-inherited allows one to estimate the recombination rate between them, and hence to also estimate the physical distance separating them. Use of multiple genetic markers allows refinement of the position of the disease locus. Since its early success in the 1980s, linkage analysis has been used to identify many monogenic disease genes. However, it meets a lot of difficulties when applied to complex diseases. There are two major problems. The first problem is related to the identification of a disease phenotype. Unlike monogenic diseases that are relatively easy to be diagnosed because of their high severity, the precise description of a complex disease is no easy task and thus a clear diagnostic standard is needed (Botstein and Risch 2003). The other major challenge arises from a consequence of a susceptibility being determined by multiple loci. As a consequence, each disease-susceptible allele may only possess modest relative risk. Risch and Merikangas have estimated that in order to achieve statistically significant results for a complex disease, the number of family pedigrees needed is impractically large (Risch and Merikangas 1996).

Association Studies

The choice for complex disease

Association studies, also referred to as case-control studies, measure the association of a genetic marker, usually a SNP, with a disease trait in a population. A marker is assumed to be associated with a disease if it has a significantly higher frequency in the patient population than the control population. Currently, it is believed that association studies offer a more practical approach for common human disease than linkage analysis. A large population sample is easier to get than a large number of suitable families for linkage analysis, which would be needed to obtain sufficient statistical power. A population can be regarded as a multigenerational family descended from a single or a few founders. Inside such a big family, all but a few most recent generations are missing. Numerous historical recombination events are included, so that only markers very close to a disease mutation will be in LD (Linkage disequilibrium) with the disease mutation. While this feature allows high precision mapping of disease genes, it does require a large number of markers unless the location is already approximately known. Association studies can be conducted within a list of disease candidate genes or on the whole genome (Botstein and Risch 2003).

Linkage Disequilibrium

Linkage disequilibrium (LD) is the non-random association between genetic markers in a population. On average, there is a reverse correlation between the level of LD and distance between SNPs because the likelihood of recombination between two SNPs increases with the distance between them. However, it has been found that the extent of LD varies in the human genome (Pritchard and Przeworski 2001). The human genome contains “hotspot” regions with relatively high recombination rates. Hotspots are separated by relatively large haplotype blocks in which there are low recombination rates. A Haplotype is defined as a set of strongly associated alleles, usually SNPs, inside such a block. There are typically a small number of common haplotypes at each genomic region in a population. Usually presence and absence of a few SNPs can be used to determine which haplotype is present. These are called haplotype tag SNPs. Using tag SNPs rather than all the SNPs will decrease the number of markers required for association studies (Gabriel et al. 2002).

Haplotype blocks are heterogeneous among different populations (Gabriel et al. 2002). European and Asian populations are relatively more homogenous, with larger blocks. African populations are more heterogeneous, and contain relatively small haplotype blocks. Tishkoff et al. (Tishkoff and Verrelli 2003) argued that this difference may be due to several possible factors: African populations are older, African populations have a larger effective population size and non-African populations experienced a bottleneck event. It has been proposed that European and Asian populations first be used to map a disease trait to a certain chromosomal region, and then African populations with smaller haplotype blocks can be used for fine-scale mapping (Tishkoff and Verrelli 2003).

Bioinformatics

So far, association studies have not delivered many successes in mapping complex disease, because of many challenges, including the possibility that complex diseases are caused by many rare variants and other statistical problems. On the other hand, the knowledge of the human genome sequence (Lander et al. 2001; Venter et al. 2001) and a large number of SNPs opens the way for the development of a detailed understanding of the mechanisms by which genetic variation results in phenotype variation. In particular, it should now be possible to identify the contribution of SNPs to human disease and thus provide a set of prioritized SNPs for association studies. Computational analysis of SNPs can be used as complementary data to confirm positive association study results and to identify causative mutations, which should be especially helpful when a disease gene is inside a large LD block. Informatics also has a significant role to play in relating the effect of an individual SNP in a protein to the gene-gene network environment and hence to the phenotypic impact, as discussed later.

Computational Modeling of SNPs: Overview

Mis-sense variants are the most frequent known monogenetic disease mutation. As a consequence, many computational methods have been developed to model the impact of mis-sense variants on protein function *in vivo*. All of these methods require some formal training and testing to assess their ability to distinguish between disease and non-disease data. A variety of approaches have been used to collect deleterious mis-sense variants and non-deleterious mis-sense variants for training and testing purposes. Wang and Moult (Wang and Moult 2001) used a set of known human monogenic disease mis-sense variants in the Human Gene Mutation Database as the deleterious data set. Sunyaev et al. (Ramensky et al. 2002; Sunyaev et al. 2001) used disease variants annotated in the Swiss-Prot database (Boeckmann et al. 2003) as the deleterious data set. Others (Chasman and Adams 2001; Krishnan and Westhead 2003; Ng and Henikoff 2003) used mis-sense mutants that affect the phenotype in mutagenesis experiments on Lac repressor and T4 lysozyme as the deleterious data set. Two different methods have been used to collect non-deleterious variants. Sunyaev et al. (Ramensky et al. 2002; Sunyaev et al. 2001) used non-synonymous base differences between human proteins and closely related proteins in other mammals. Other groups (Chasman and Adams 2001; Krishnan and Westhead 2003; Ng and Henikoff 2003) used the mis-sense mutants that do not show phenotype effects in mutagenesis data for Lac repressor and T4 lysozyme as the non-deleterious data set.

Two principal strategies have been developed for identifying which mis-sense base changes are most likely to be causative of disease. The most common approach makes use of the fact that the more critical a position in a protein sequence is to viability, the more restricted are the residue types accepted there. A number of different methods for assessing the significance of amino acid conservation have been developed (Chasman and Adams 2001; Krishnan and Westhead 2003; Ng and Henikoff 2003; Ramensky et al. 2002; Sunyaev et al. 2001). Methods that utilize sequence conservation have the advantage of including all kinds of impact on protein viability. Also, these methods can be used with any human protein for which a suitable set of sequence relatives is known, and so have wide applicability. The approach has the disadvantage that it provides no direct insight into the underlying mechanism. The second strategy is to make use of knowledge of protein structure and function. For instance, recognizing that a change occurs in a key catalytic residue, or one involved in ligand binding, or a target for post-translational modification.

Wang and Moult (Wang and Moult 2001) used a structure-based model to identify amino acid substitutions likely to significantly affect protein stability as well as other contributions to function. Stability impact was assessed using a set of simple rules based on changes in hydrophobic burial, backbone strain, overpacking, and electrostatic interactions. Other groups have combined sequence and structure strategies to varying degrees. Sunyaev (Ramensky et al. 2002; Sunyaev et al. 2001) predicted the effect of mis-sense mutations using empirically derived rules which make use of a variety of data, such as functional information, hydrophobic propensity, side-chain volume change and transmembrane location (Ng et al. 2000), together with sequence information. In Chasman's (Chasman and Adams 2001) method, ANOVA (Analysis of variance) and principal component analysis were applied to a series of features that capture aspects of structural and sequence context. Features showing strong discrimination between mutations affecting or not affecting the phenotype, such as the relative residue temperature factor, relative surface accessibility, relative phylogenetic entropy (sequence conservation in the protein family) and burial of charge, were selected. A probability model was then constructed based on the selected features, and used to estimate the likelihood that a given mutation will affect function. A similar probability approach has also been used to include function effects (Lau and Chasman 2004). Krishnan and Westhead (Krishnan and Westhead 2003) used two machine learning methods, a decision tree and a support vector machine, to predict the impact of single amino acid changes based on a set of structural (secondary structure and surface accessibility) and sequence attributes, such as sequence conservation score calculated using ScoreCons (Valdar and Thornton 2001). Secondary structure and surface accessibility data were taken from the HSSP database (Dodge et al. 1998) or predicted using PHD (Przybylski and Rost 2002).

Computational Modeling of SNPs: Hypotheses

In this project, we have developed both a structure based model and a sequence based model. The central hypothesis of the structure-based model is that moderate loss of stability of the folded state of a protein molecule is frequently associated with monogenic disease. In principle, the stability could be assessed with numerical free energy calculations. However, at present, such calculations are not reliable enough to provide useful answers (Mark and van Gunsteren 1994). Knowledge-based methods, dividing stability into a set of factors such as electrostatic interaction and hydrophobic burial, provide an alternative approach. In the present work, a knowledge based method has been used to estimate whether or not an amino acid substitution reduces protein structure stability significantly. 15 factors that are related to the free energy of the folded state of protein are used. They are described more fully later. A machine learning technique (a support vector machine, SVM (Vapnik 1995), see the following section) is used to distinguish deleterious mis-sense variants and non-deleterious variants using these 15 factors.

The underlying hypothesis for the sequence based model is that deleterious mutations would be removed during evolution and thus critical amino acids tend to be conserved across species. Five features that capture the relative sequence conservation at each position in a multiple-species sequence alignment are used as basis for SVM to identify deleterious mis-sense variants. Those mutations at a conserved position tend to be the most deleterious.

SNPs3D: a resource for analysis of SNP, identification of candidate genes and construction of gene relationship networks

Some SNP analysis methods form the basis of tools that are available through web servers. Facilities range from tools to visualize SNPs in their three dimensional context, such as MutDB (http://www.mutdb.org) (Dantzer et al. 2005; Mooney and Altman 2003), TopoSNP (http://gila-fw.bioengr.uic.edu/snp/toposnp) (Stitziel et al. 2004; Stitziel et al. 2003), SAAP (http://www.bioinf.org.uk/saap/) (Cavallo and Martin 2005), to detailed analysis of the molecular effects of mis-sense SNPs. For example, SNPeffect (http://snpeffect.vib.be/) provides a comprehensive analysis of mis-sense SNPs at the protein level (Reumers et al. 2005) including stability analysis using FOLD-X (Guerois et al. 2002), and other functional analysis; PolyPhen (http://www.bork.embl-heidelberg.de/PolyPhen) models SNP effects with both structure and sequence information (Ramensky et al. 2002); SIFT (http://blocks.fhcrc.org/sift/SIFT.html) provides sequence analysis of mis-sense SNPs (Ng and Henikoff 2003).

In the present work, we have used the structure and sequence based models to identify deleterious SNPs in the current version of dbSNP (Build 124), and a publicly available website, SNPs3D, has been developed to provide easy access to our analysis for the scientific community.

To maximize use of data, it is necessary to put SNP analysis into the pathway context. A number of projects, including the Ingenuity Pathway database (http://www.ingenuity.com) and the Protein Reference Database (Peri et al. 2004), (http://www.hprd.org), are developing mammalian pathway descriptions by means of manual curation of the literature. Although these databases provide rather precise data, the human-curation process makes development slow. This problem is becoming more serious as the size of the relevant literature increases. Protein interaction networks have also been built automatically (Giot et al. 2003; Lee et al. 2004) (Li et al. 2004; Tong et al. 2004), using probability models to integrate data from high throughput experiments such as yeast-2-hybrid (Fields and Song 1989; Phizicky et al. 2003) and TAP pull-downs (Jansen et al. 2003).

In SNPs3D, a network of gene-gene interactions is derived from the literature. A variety of computational methods are being developed to automatically extract information from the literature. These methods range from simple technologies which process at the word level and require only a limited linguistic context (Stapley and Benoit 2000) to state of the art technologies such as natural language processing (NLP), that handle more complex relations across sentences (Daraselia et al. 2004). So far, these methods have not been used extensively in generally available gene-disease interfaces.

We make use of simple text mining techniques. Each gene or disease is treated as a concept. Words and terms are extracted from relevant PubMed abstracts ordered by their relevance to the concept. The overlap of the keyword profiles between a pair of genes is used to build a gene relationship network. Profiles are also used to provide a list of candidate genes for diseases. A Java interface has been developed to display gene-gene relationship. The Java interface also allows access to a range of relevant information including pathways (Kanehisa et al. 2004), mRNA expression profiles (Su et al. 2002), mouse knockout (http://www.bioscience.org

/knockout/knochome.htm), disease-gene relationship databases (Hamosh et al. 2005; Stenson et al. 2003) and the underlying literature.

Technology overview

Classification

The work described in this thesis makes use of a machine learning technique, the support vector machine (SVM), to classify mis-sense mutations as deleterious or non-deleterious to protein function *in vivo*. The support vector machine is a computational technique for data classification.

SVMs are one of a large number of machine learning techniques, including Decision Trees (Markey et al. 2003; Sachs 2003), Neural Networks (Bidiwala and Pittman 2004; Gromiha et al. 2004), Bayesian Networks (Li and Chan 2004; Nariai et al. 2004). These techniques have been used in solving many scientific and commercial classification problems. SVMs were introduced in 1995 and have spread into many fields, because they are easy to handle and are usually among the top-ranking algorithms (Rost and Eyrich 2001). They have been used successfully to solve a number of biological questions (Bhasin et al. 2005; Hua and Sun 2001; Mitsumori et al. 2005; Tsirigos and Rigoutsos 2005; Zhao et al. 2005).

In the present application, we wish to classify mis-sense mutations as deleterious or non-deleterious on the basis of a set of features. The features used for the structure based model are 15 stability parameters, such as overpacking, electrostatic interactions and hydrophobic effects. Five features related to sequence conservation were used for the sequence model. Figure 1-1 shows a simple example for a two feature system.

A central issue in all data classification is over-fitting. When a training set is presented to a learning algorithm, the algorithm usually tries to find a model which correctly classifies as many of the objects as possible. A very complicated model may perfectly classify the objects in the training set, but may poorly classify new observations because the model is too specific. This phenomenon is known as over-fitting. Cross-validation is used to objectively test the usefulness of a model, by training on one dataset and testing on a different one. In this project, disease causing variations and non-deleterious variations are separated into two groups: one group functions as a training set, and the other is used to validate each model.

Comparative modeling

A protein structure is required in order to evaluate the stability effect of a mis-sense variant. Currently, experimentally-solved 3D structures are only available for a small fraction of human proteins. The number of proteins for which structural information is available can be increased significantly by modeling. Comparative modeling makes use the fact that when two proteins have similar sequences, indicating a common evolutionary origin, their 3D structures will also be similar. As the sequence identity between two proteins decreases, their structural similarity also decreases, so that the most reliable models are based on a high level of sequence relationship. A protein whose structure is to be built is called a target. A protein with a sequence similar to that of a target and with a solved structure is called a template. A comparative model is constructed simply by coping structural coordinates of

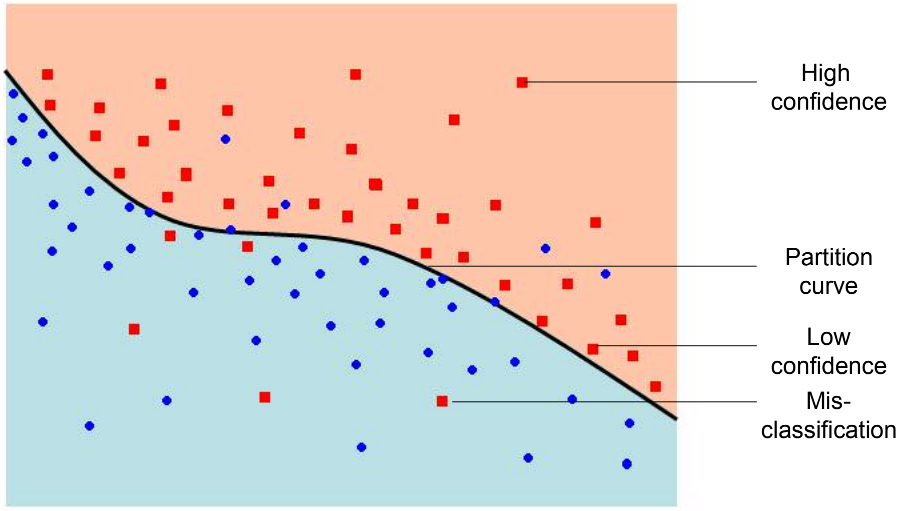


Figure 1-1. Illustration of a Support Vector Machine.

There are two classes of data represented by squares and round points. Each data point has two features, defining the X and Y coordinates. The support vector machine selects a partition in the space (in this case the curved line) that separates the two data sets as far as possible. Such a partition is seldom perfect and some points will be mis-classified. In general, the further from the partition surface a point lies, the higher the confidence. In this thesis, the data are mis-sense mutations, which are classified as deleterious or non-deleterious to protein function *in vivo*.

related regions of a template to create a target model and changing the side-chains types and coordinates as necessary. According to the results of CASP (Critical Assessment of Structure Prediction), models based on a sequence identity around 30% have approximately 1.5 Å root mean square (RMS) error for main chain atoms (Baker and Sali 2001). As we shall see later, this is not quite good enough for our purpose, and a 40% threshold is used. Figure 1-2 illustrates the comparative modeling process.

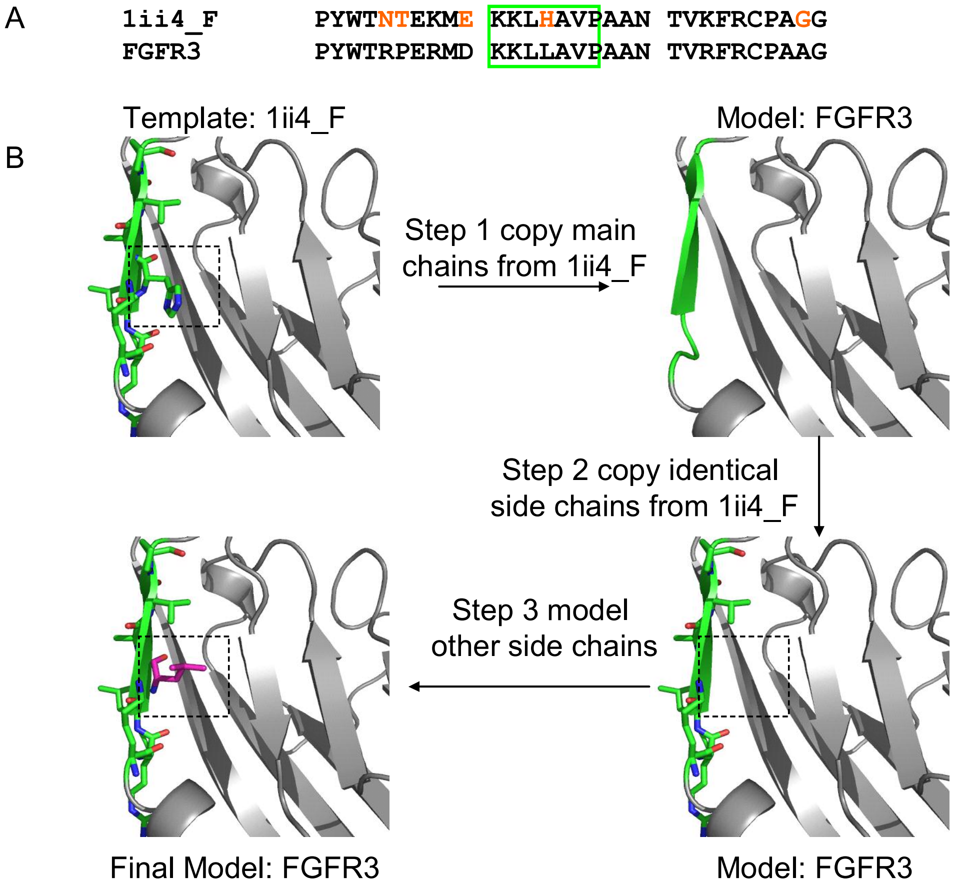

A
1ii4_F PYWTNTEKME KKLHAVPAAN TVKFRCPAGG
FGFR3 PYWTRPERMD KKLLAVPAAN TVRFRCPAAG
Template: 1ii4_F
Model: FGFR3
B
Step 1 copy main chains from 1ii4_F
Step 2 copy identical side chains from 1ii4_F
Step 3 model other side chains
Final Model: FGFR3
Model: FGFR3

Figure 1-2. The process of comparative modeling.

A. Part of the sequence alignment between the protein of a gene FGFR3 and the closest available template structure, PDB (Deshpande et al. 2005) code 1ii4_F. The boxed region is used to illustrate the detailed procedure of comparative modeling. B. Comparative modeling, (the structure represented by the green area in Figure B corresponds to residues inside the rectangle in Figure A). Step 1. Main-chain coordinates are copied to the target from the template based on the sequence alignment, for example, the main-chain coordinates of Histidine on the template are copied to Leucine on the target. Step 2. Side-chain coordinates are copied to the target from the corresponding identical template residues. Step 3. The remaining side-chains are modeled, for example the purple for Leucine.

Chapter 2: Loss of Protein Structure Stability as a Major Causative Factor in Monogenic Disease

Introduction

The central hypothesis of the present work is that moderate loss of stability of the folded state of a protein molecule is frequently associated with monogenic disease. To investigate this, we must identify significant changes in the free energy difference between the folded and unfolded states of a protein molecule resulting from an amino acid substitution. A theoretically rigorous approach would be to use an appropriate integration of the energy change as one amino acid is morphed into another in the context of the protein structure. These free energy perturbation techniques (Beveridge and DiCapua 1989) have been incorporated in a number of the more widely used molecular dynamics software packages. Issues of conformational sampling, appropriate representation of the unfolded state and force field accuracy have generally resulted in poor accuracy (Mark and van Gunsteren 1994). Recent results show encouraging improvement, but require care and method optimization in each case (Pan and Daggett 2001), restricting large scale application. Force field deficiencies may be reduced by parameterizing using free energy differences obtained from site directed mutagenesis experiments (Guerois et al. 2002). The resulting model is effective at predicting this type of stability change.

We have developed a knowledge based method which estimates whether or not an amino acid substitution reduces protein structure stability sufficiently to be potentially causative of monogenic disease, As in the earlier work (Wang and Moult 2001), we make use of the extensive literature on the effect of amino acid substitutions on protein stability, as well as knowledge of the underlying factors affecting the free energy of the folded state. We identify a set of 15 such factors that may contribute to a free energy difference, through changes in interaction energy between amino acids, effects on the entropy of the system, and the local rigidity of the structure. A machine learning technique (a support vector machine (Vapnik 1995)) is used to partition the 15 dimensional space representing these factors into two volumes, in such a way that, as far as possible, disease causing mutations fall in one volume and non-disease causing ones in the other. Any new mutation may then be assigned a position in this space. Mutations falling in one volume are predicted to significantly decrease protein stability, and thus to be potentially disease causing. Those falling in the other volume are considered non-disease causing. Distance from the volume partitioning surface provides an approximate measure of confidence in the assignments, as illustrated in figure 1-1.

The model is trained on a set of mis-sense mutations that cause monogenic disease, extracted from the Human Gene Mutation Database (HGMD (Stenson et al. 2003)). A control set of residue substitutions not contributing to disease susceptibility was based on inter-species differences (Sunyaev et al. 2001). Stability effects are analyzed using available experimental structures of human proteins, or reliable comparative models. Jack-knifed testing shows that this model does differentiate between disease and non-disease mutations, validating the hypothesis that stability effects play a major and quite general role in monogenic disease.

Results

Selection of Data for Analysis

As described in the methods, 10,263 disease causing mutations in 731 proteins were extracted from the HGMD (Stenson et al. 2003). Appropriate structure information was available for 37% (3768 in 243 proteins) of these mutants, forming the disease set. 346 of the HGMD proteins had close orthologs in other species. The corresponding 16,682 inter-ortholog residue differences provided a set of non-disease variants. 14% (2309 in 153 proteins) of the inter-species variants had appropriate structure information, and formed the control set.

Analysis of Factors Likely to Affect Protein Stability

Eleven contributions to the energy and entropy of proteins stability are considered. There are four classes of electrostatic interaction: reduction of charge-charge, charge-polar or polar-polar energy, or introduction of electrostatic repulsion; three solvation effects: burying of charge or polar groups, and reduction in non-polar area buried on folding; and two terms representing steric strain: backbone strain and overpacking. The other two contributions considered are cavity formation (affecting van der Waals energy), and loss of a disulfide bridge. Figure 2-1 shows examples of each of these, with the corresponding disease outcome. The crystallographic temperature factor and surface accessibility of mutated residues are also considered.

Figure 2-2A shows the distribution of each of these effects in the disease and non-disease data sets. (Criteria used are described in the Methods section). The red bar shows the fraction of all disease data points classified as disease, and the green

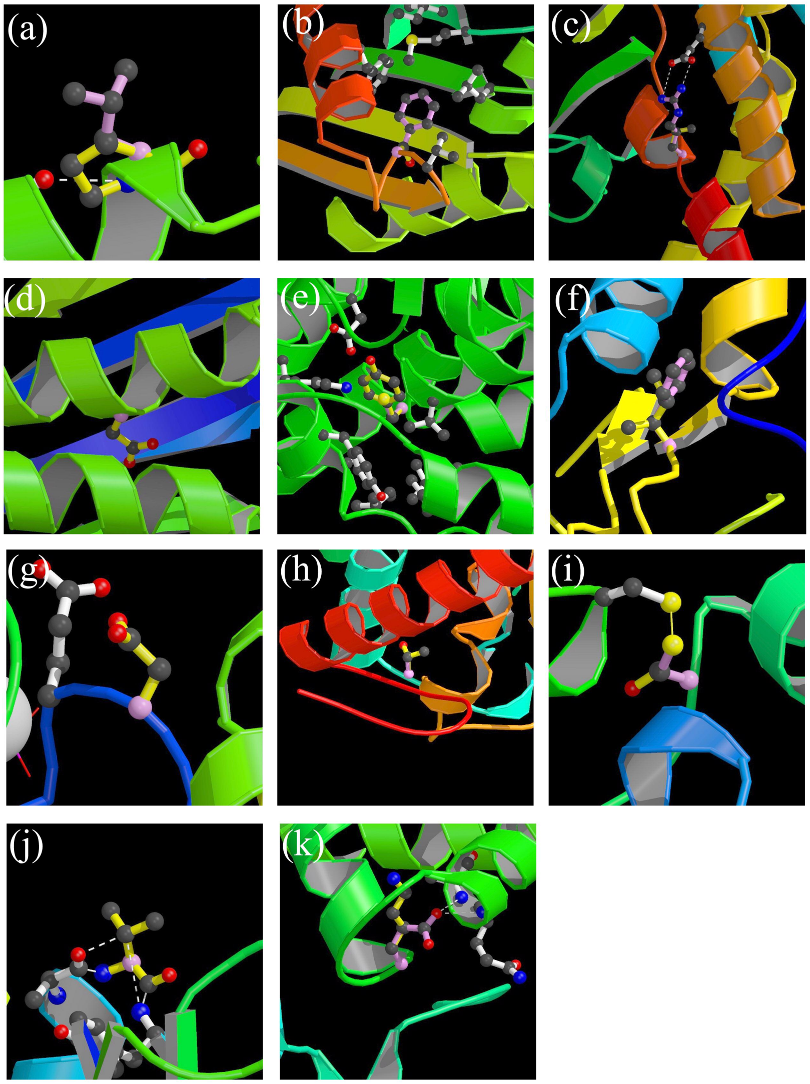


Figure 2-1. Examples of disease caused by structure destabilizing factors.

For each case, bonds of wild type side chains are shown purple, and bonds of the mutant side chains are yellow. Atoms are colored by type. In a number of cases, more than one factor is involved. The selected one is judged to be the most significant. The full model considers all factors together. Disease associations are taken from the NCBI Refseq database.

(a) Loss of polar-polar interactions. L226P in galactose-1-phosphate uridylyltransferase (GALT, PDB code 1HXP_B), causing galactosemia. This mutant introduces a proline into an alpha helix, resulting in the loss of a main chain hydrogen bond, as well as loss of hydrophobic interactions of the side chain.

(b) Loss of hydrophobic interactions. F234S in GTP cyclohydrolase (GCH1, 1IR8_I), causing dopamine-responsive dystonia. A large buried non-polar side chain is replaced by a small polar one, reducing the burial of non-polar area on folding. A cavity is also created, and there is a small gain in polar-polar energy.

(c) Loss of a salt bridge. R382L in isovaleryl Coenzyme A dehydrogenase (IVD, 1IVH_C), causing isovaleric acidemia. R382 forms a salt bridge (charge-charge interaction) in the wild type protein, lost in this mutant.

(d) Buried charge. G60D in aspartylglucosaminidase (AGA, 1APY_A), causing aspartylglycosaminuria. G60D introduces a charge group into the interior of the protein. It also causes over-packing.

(e) Overpacking. C91Y acyl-Coenzyme A dehydrogenase (ACADM, 1EGE_C), causing ACADM hereditary deficiency. C91Y introduces a bulky side chain into the interior of the protein, resulting in substantial overpacking.

(f) Cavity formation. F411I in glucocerebrosidase (GBA, 1OGS_A), causing Gaucher's Disease. F411I replaces a large buried non-polar side chain with a smaller one, creating an internal cavity. There is also a loss of hydrophobic interaction.

(g) Electrostatic repulsion. G38D in guanine nucleotide binding protein (GNAT1, 1TAG), causing night blindness. Introduction of the aspartic acid side chain results in an unavoidable electrostatic repulsion with another aspartic acid. There is also limited overpacking.

(h) Buried polar group. A543T in Hexosaminidase B (HEXB, 1O7A_D, causing Sandhoff Disease. Here a hydroxyl group is introduced in a buried non-polar environment. There is also minor overpacking.

(i) Breaking of a disulfide bond. C163S in aspartylglucosaminidase (AGA, 1APY_A), causing aspartylglycosaminuria. C163S replaces one component of a disulfide bond.

(j) Backbone strain. G137V in arylsulfatase B (ARSB, 1FSU), causing Maroteaux-Lamy syndrome. G137V introduces a side chain onto a glycine residue with backbone dihedral angles unsuitable for other residue types.

(k) Loss of charge-polar interaction. E167K in uroporphyrinogen decarboxylase (UROD, 1R3Q_A), causing familial porphyria cutanea tarda and hepatoerythropoetic porphyria. E167 forms charge – polar interactions with two main chain N-H groups, providing a helix cap. The mutation removes these interactions.

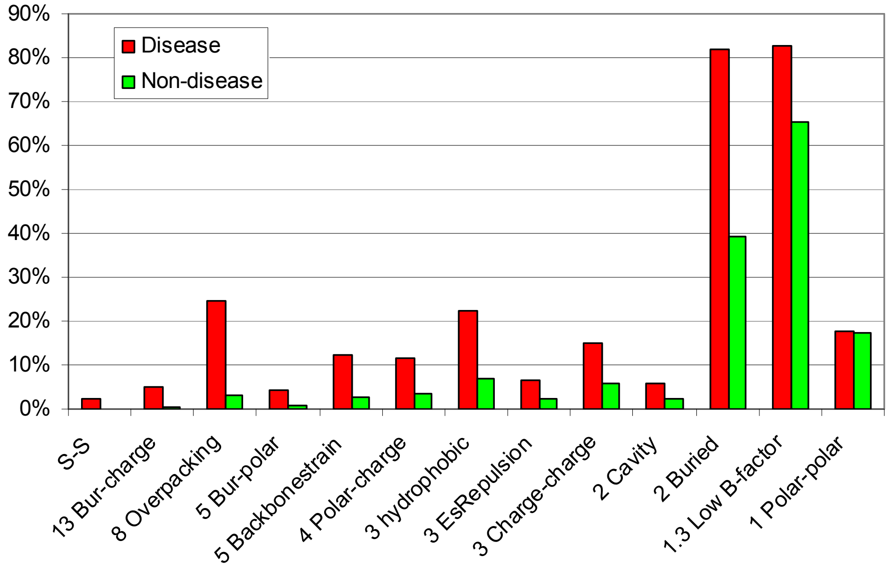


Figure 2-2A. Partitioning of each stability factor between the disease and non-disease data sets.

The red bars show the fraction of disease variants covered by the corresponding factor, and the green bars show the fraction of non-disease variants covered. An ideal factor has high coverage of the disease set, and no examples in the non-disease set. Factors are ordered by the discriminatory power (ratio of disease to non-disease coverage), best discriminators to the left. The discriminatory power of each factor is included in the bar labels. The ratio ranges from infinite for breaking a disulfide bridge (no examples in the non-disease set) to 1 for polar-polar interactions (an approximately equal number of examples in the disease and non-disease sets).

bar is the fraction all non-disease points classified as disease. An ideal factor includes a large fraction of the disease points (red bars), and no non-disease points (green bar). The 11 energy and entropy factors are ordered by the ratio of the two bar heights, with the best discriminators on the left.

Discrimination power ranges from perfect for disulfide bond breakage – (the only instances are in the disease set), to none (loss of polar-polar interactions is as common in the disease set as in the control set). Coverage also varies widely, from only 3% of disease cases involving disulphide bond loss to 24% of cases involving over-packing. The last two terms capture the ability of the structure to relax to partly compensate for unfavorable energy or entropic effects. As expected, regions of lower crystallographic temperature factor contain more disease mutations than non-disease ones. Similarly, buried residues, which generally have least space to adjust to change and more other energetic restrictions, have a two fold excess of disease mutations over non-disease ones.

Greater discrimination can be achieved by taking advantage of the fact that most mutants affect more than one factor. Figure 2-2B shows some examples of discrimination using pairs of factors. For example, combining loss of a polar-polar interaction with a non-surface environment increases the ratio of disease to non-disease cases from about one to approximately three. Highest discrimination will be obtained with a method that considers all the factors affected by a mutation. For this purpose, each mutant is represented as a point in a 15 dimensional factor space. Eleven of the dimensions are the energy and entropy factors shown in figure 2-2A. One dimension is the surface accessibility of the mutated residue, relative to

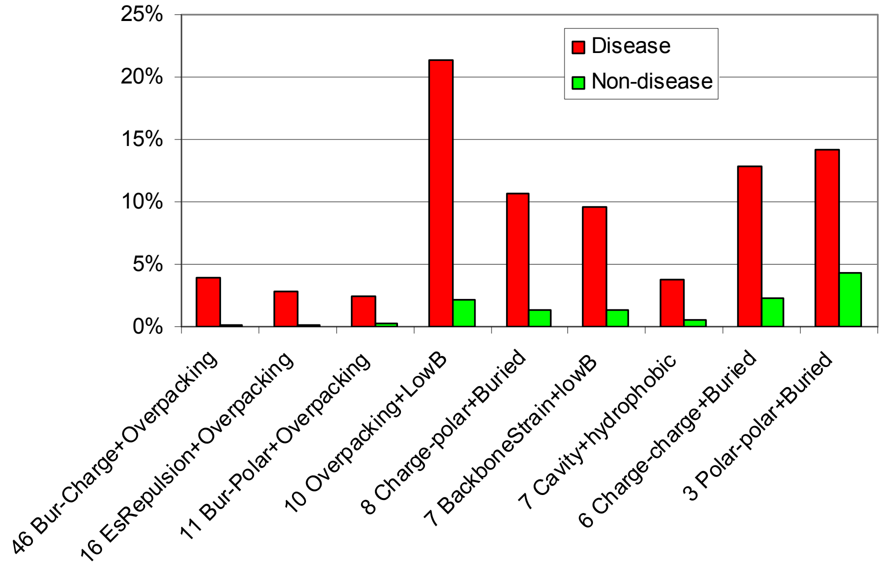


Figure 2-2B. Improvement in discrimination when two stability factors are considered together.

As in 2-2A, bars show the partitioning between disease and non-disease variants, now considering two factors at a time. Discriminatory power is considerably improved. For example, adding a non-surface requirement to loss of polar-polar interactions increase the discriminatory ratio from 1 to 3. Best discrimination is achieved when all relevant factors as considered, as in the full model.

the unfolded state. The other three are the Cα temperature factor of the mutated residues, the Z value of the temperature factor, and the standard deviation of all Cα temperature factors. (Three dimensions rather than one are used to allow for variable scaling of the experimental values). As described in Methods, a support vector machine was used to determine a surface that optimally partitions the disease and non-disease points in this space.

Accuracy of the SVM Model

Figure 2-3 summarizes the results of the model. 74% of the 3768 mis-sense mutations in the disease dataset were assigned as disease causing, and 85% of the 2309 mis-sense mutations in the non-disease dataset were classified as non-disease. For the 82% of data points more than a distance of 0.5 from the SVM partitioning surface, the prediction accuracy increases to 79% correctly identified disease data points, and 89% correctly assigned non-disease points. The 15% false positive rate arises from defects in the model. Since only stability factors are included in the model, all mutants that act through other mechanisms, such as effects on catalysis, binding and so on, are included in the 26% false negative rate. Some fraction of false negatives are mutants included in the HGMD database that do not appear to cause disease. For example, the mutant G15D in the alpha chain of Hemoglobin (HBA1) is in HGMD, but is predicted to be non-disease causing, with a confident SVM score of 0.8. The literature on this mutation (Molchanova et al. 1994) gives no indication of disease. Allowing for approximations in the model, a conservative conclusion is that substantially more than half of disease mutants operate at least partly through destabilization of the folded structure.

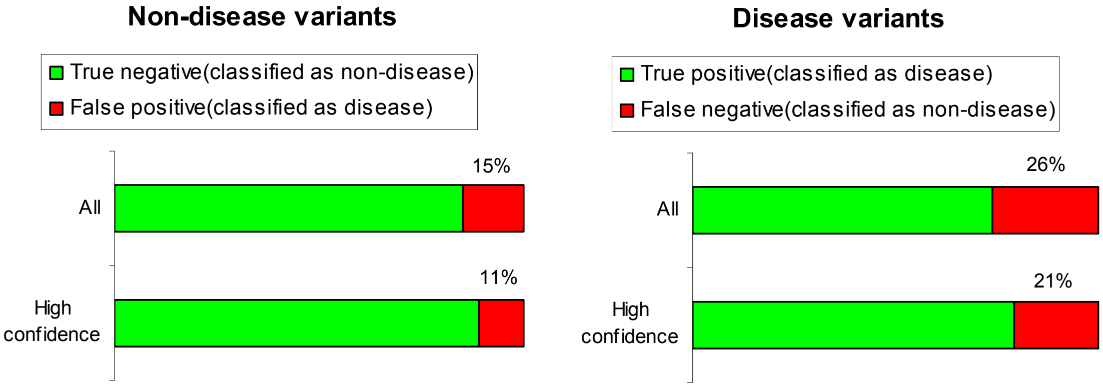


Figure 2-3. Evaluation of the Support Vector Machine model.

The left hand panel shows the fraction of disease variants correctly identified by the model in jack-knifed testing. The model is trained only to detect variants that cause disease by destabilization of the structure, so that the false negative rate of 26% includes all other causes, as well as deficiencies in the model. The bottom bar shows the result for the more confident subset of predictions (the 80% of the data with an SVM distance greater than 0.5), with a false negative rate of 21%. The right hand panel shows the same data for the non-disease data set. Here, the false positive rate (variants incorrectly assigned to disease) is 15% for the full set and 11% higher confidence classifications.

Model Evaluation using in vitro Mutagenesis Stability Data

The SVM disease model is trained entirely on disease related mutant data, containing no explicit information about stability. Evaluation of the model's performance against *in vitro* mutagenesis free energy data provides an independent test of the hypothesis that disease is strongly coupled with structure destabilization. We would expect that there should be a strong correlation between a potential disease outcome and the change in the free energy difference between the folded and unfolded states.

As described in 'Methods', we have run the disease trained prediction model against a set of 581 of these *in vitro* stability data, from four proteins (Table 2-1). Figure 2-4 shows the relationship between the change in free energy and the fraction of mutations that would be predicted to have a disease outcome. For mutants that stabilize or mildly destabilize the folded state (up to 1 Kcal/mole) the fraction of potential disease causing residues is close to the false positive rate of the model (16%). As the change in free energy increases, so does the fraction of potential disease-causing mutations, reaching 90% in the 3 - 4 kcal/mol range, and 100% above 4 kcal/mol. These results confirm that the model is detecting destabilizing effects on structure. The observation that most potential disease-classified mutations destabilize the folded state by about 2 to 3 Kcal/mol suggests that real disease causing mutations will be in this range. That conclusion is supported by the fact that the distribution of SVM scores for mutants that destabilize by

Protein and PDB structure	Structure class	Number of Mutations (Total 581)
Acyl-coenzyme A binding protein (2abd)	All alpha	30 (Kragelund et al. 1999)
fk506 binding protein (1fkj)	Alpha and beta	34 (Fulton et al. 1999; Main et al. 1998)
Barnase (1bni)	Alpha and beta	87 (Serrano et al. 1992a; Serrano et al. 1992b; Serrano et al. 1992d)
Staphylococcal nuclease (1stn)	All beta	430 (Byrne et al. 1995; Green et al. 1992; Green and Shortle 1993; Holder et al. 2001; Meeker et al. 1996; Schwehm et al. 1998; Shortle et al. 1990; Stites et al. 1994)

Table 2-1. *In vitro* mutagenesis data from four proteins, used to test the SVM model. Structure class is taken from SCOP (Andreeva et al. 2004)

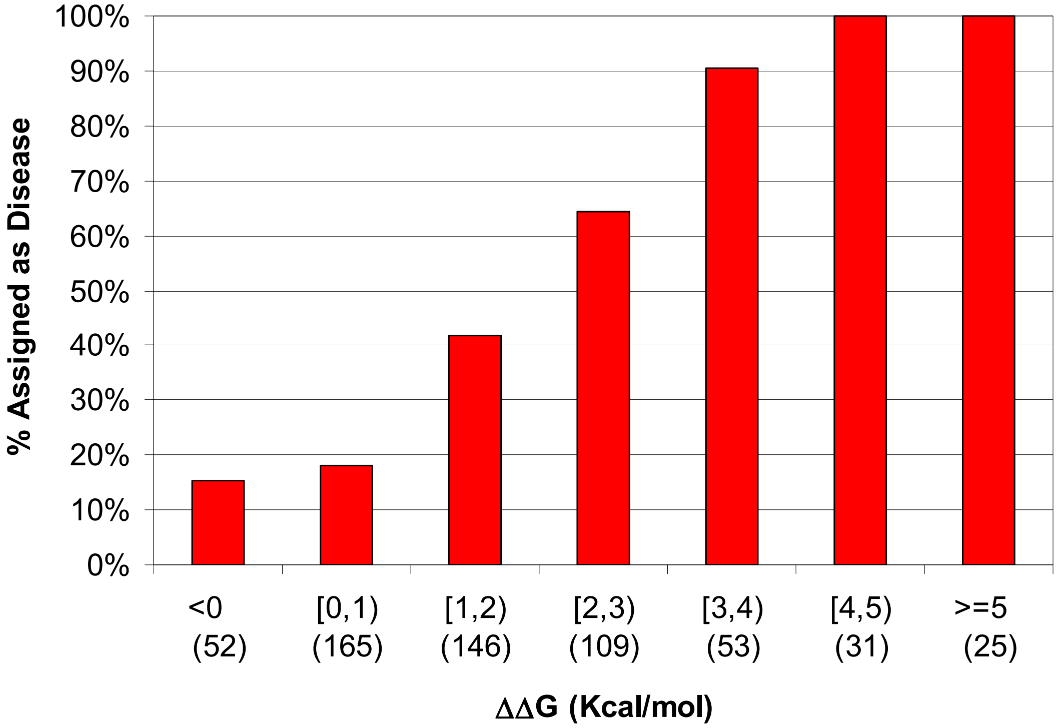


Figure 2-4. Application of the Disease/Stability model to *in vitro* site directed mutagenesis data.

The plot shows the fraction of mutants classified as consistent with disease, as a function of the free energy difference between the folded and unfolded states. For stabilizing and weakly destabilizing mutants, the disease compatible fraction is similar to the false positive rate of the model. Above 3 Kcal/mol of destabilization, 90% of mutants are classified as disease compatible. The results suggest that a typical disease causing mutant destabilizes the folded state by 2 – 3 Kcal/mol.

more than 2 Kcal/mol is similar to that of the disease causing mutants (means of -0.88 and -1.00, medians of -0.68 and -0.60, respectively).

It is informative to examine the outliers in this distribution. Five (L108I, L36V, L37V and A132G in staphylococcal nuclease and S92A in barnase) of the 53 mutants that decrease stability by 3 to 4 Kcal/mol are predicted not to be consistent with disease. The two L -> V mutants differ by one methyl group, and both result in a slight loss of hydrophobic burial. There are 24 L -> V mutants in the disease dataset and 37 cases in the non-disease dataset, suggesting that this class of mutant is finely balanced between disease and non-disease causing, and subtle effects tip the balance. Consistent with this, the SVM gives a low confidence score (0.14 and 0.13) for these two outliers. L108I creates no change of volume or overall hydrophobicity, so it is surprising that it is so destabilizing. There are 25 such mutations in the non-disease dataset, and only four in the disease set, suggesting this high level of destabilization is unusual. The SVM score is also in the less confident range (0.3). The authors of the experimental study (Holder et al. 2001) suggest loss of highly optimal van der Waals packing is primarily responsible for the large effect. The remaining two mutations, A132G and S92A, are both predicted to be non-disease causing with relatively high confidence (SVM scores 0.70 and 0.89). For A132G, there is a minor loss of hydrophobic burial. There are 36 cases of A -> G mutations in the non-disease set and only 11 cases in the disease dataset. For S92A, the model identified the loss of a hydrogen bond and a slight gain of hydrophobic burial. Serrano and colleagues (Serrano et al. 1992a; Serrano et al. 1992b) note that this residue is the first residue in a beta turn between two strands. The hydroxyl group is buried, and makes two hydrogen bonds, suggesting it may be involved in unusually strong interactions. There are 77 cases of S -> A mutants in the non-disease set and only two in the disease set, indicating that such strong polar electrostatic interactions are unusual.

Eight of the 52 mutants that increase protein stability are predicted to be consistent with disease. All but one are in Staphylococcal Nuclease. All increase stability by less than 1 Kcal/mol. For three cases: N138G, S128A and H124F, the SVM returns a low confidence score. In none of the other cases is it clear why there is disagreement with experiment. For D21A and D21G, there is a predicted loss of charge-charge and charge-polar interactions. For D21A and D21G, there is a predicted loss of charge-charge and charge-polar interactions. The distribution of these two mutations between the disease and non-disease datasets are 8/8 and 57/11 respectively. T41I is predicted to result in a large gain of hydrophobic burial, offset by the loss of a charge-polar and polar-polar interactions in a buried environment. There are 41 cases of T -> I mutations in the disease dataset and 18 cases in non-disease dataset, most with a predicted large gain hydrophobic burial and decreased electrostatic interactions. G50A is predicted to result in backbone strain. It is probable the structure is able to relax to accommodate the change in backbone angles. The temperature factor is moderately high, supporting this possibility. The eighth mutant, N58D, is in barnase. There is a predicted loss of polar-polar interaction and a slight gain of charge-polar interaction.

Alternative Test Sets

This work uses disease and non-disease related data for training and testing. Others (Chasman and Adams 2001; Ng and Henikoff 2003)' (Krishnan and Westhead

2003) have used data on the phenotypic impact of single residue mutants in a bacterial and a phage protein. We have investigated the relationship between our assignment of disease potential and phenotypic impact in these mutagenesis sets. The data are a set of about 4000 mutants of the E.coli lac repressor (Markiewicz et al. 1994) and a set of about 2000 mutants of phage T4 lysozyme (Rennell et al. 1991). A total of 1,987 mutations in T4 lysozyme and 3,291 mutations in lac repressor can be modeled on the corresponding protein structures (PDB entries 1lbh and 7lzm respectively). Each data set was partitioned into groups based on the phenotype annotations in the literature. For lac repressor, these annotations are: "+" (wild-type phenotype, 200 fold repression of beta-galactosidase activity, but in practice some times only 8 -10% of this); "+s" (wild-type phenotype under certain conditions, including temperature sensitive mutations); "+-" (20 -200 fold galactosidase glactosidase repression); "-+" (4 – 20 fold); and "-" (less than 4 fold repression). For T4 lysozyme, the groups are: "++" (wild-type phenotype – plaque size similar to control); "+" (signifcanlty smaller plaques); "+/-" (Similar in size to "+', but hazy morphology); and "-" (no plaques produced).

The HGMD trained SVM model was used to assign potential disease mutants in each of the phenotype categories. Figure 2-5 shows the results. For both proteins, a high fraction of the mutants in the most severe class of phenotype impact are assigned as disease like (~90% for Lac repressor and ~100% for T4 lysozyme). However, for

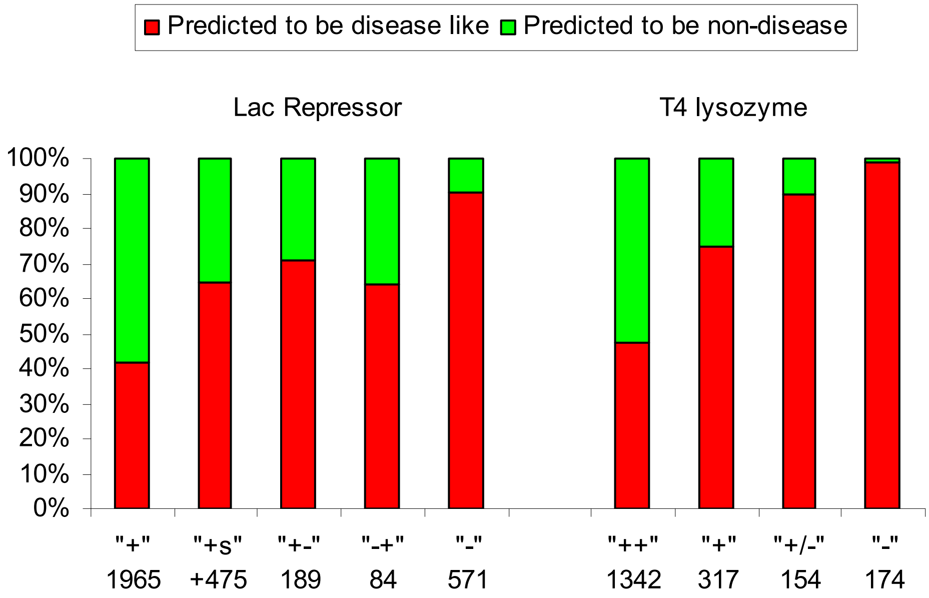


Figure 2-5. Application of the Disease/Stability model to mutants of Lac repressor and T4 lyzozyme.

Symbols below the bars indicate the extent of phenotypic impact for that set of mutants, from '+' for the most activity to '-' for none. Red regions of the bars show the fraction of mutants in each category found to be compatible with disease. As expected, a high fraction of the low activity mutants are assigned as compatible with disease, but a significant fraction of the maximum activity ones are also so classified. This result is consistent with the fact that a low % of activity is sufficient for a '+' classification for both proteins. Numbers below each column show the number of mutants in that category.

both proteins, about 40% of ‘wild type’ mutants are also assigned as consistent with disease. The probable explanation is that a rather low level of enzyme activity is needed for a wild type classification: For T4 lyzozyme, as little of 4% residual enzyme activity may be classified as wild type (Rennell et al. 1991), and for Lac repressor, 10% activity is some times sufficient (Markiewicz et al. 1994). Such low levels of monogenic disease protein activity would likely usually result in disease.

Functional Analysis of Single Residue Mutations

An advantage of the structure/stability model is that it provides mechanistic insight into why a mutant has a deleterious effect on protein function. In principle, functional roles, such as ligand binding and catalysis, may also be assigned to particular residues, and so allow more general mechanism based analysis. As described in Methods, we have investigated this possibility using SwissProt functional annotation and experimentally observed ligand binding. Figure 2-6 shows the results. For the disease set, an additional 1.6 % of the mutants that were false negatives in the stability model are annotated as functionally important. Seven percent of the stability related mutants are also assigned a functional role. These low values probably reflect the incomplete assignment of function. Inclusion of these in the model would reduce the false negative rate by 1.6%. However, in the non-disease set, an additional 2.1% of mutants are assigned a functional role, leading to an increase in the fraction of false positives. Thus, we conclude that, at present, residue function annotation is too unreliable and incomplete to be useful.

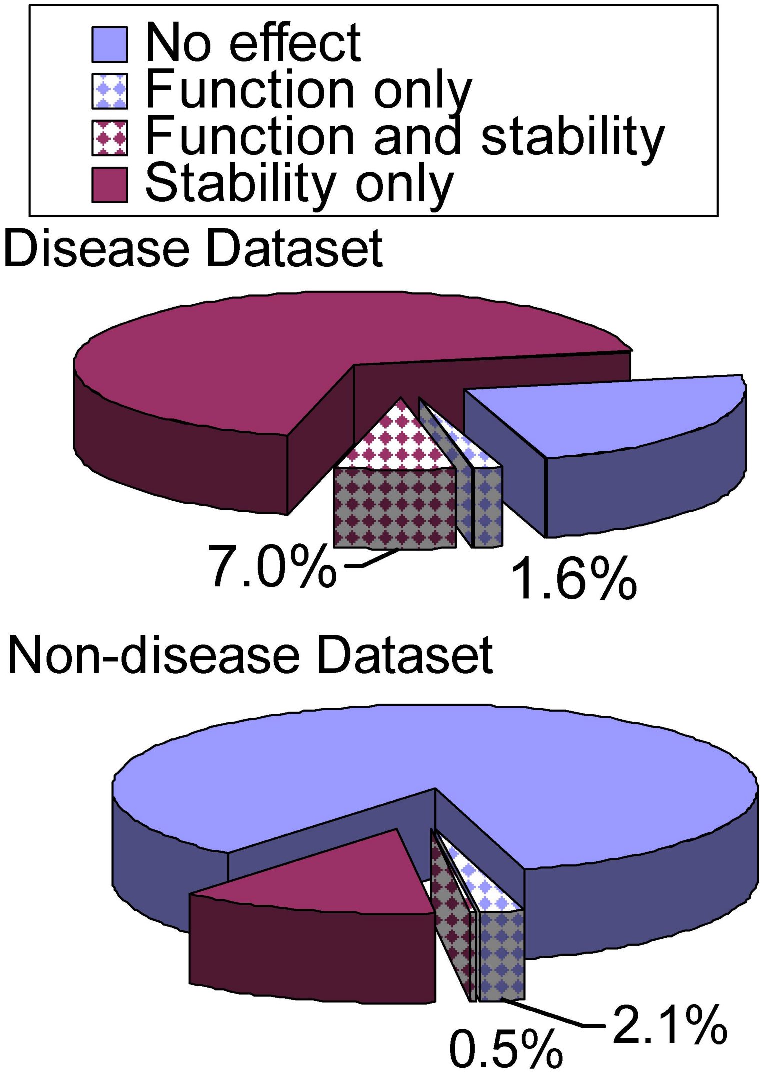


Figure 2-6. Distribution of direct functional effects of variants in the Disease and Non-disease data sets.

Residue function was assigned from Swiss Prot annotation and on the basis of contacts with bound ligands. 7% of stability variants also have a known functional role, and only an additional 1.6% of false negatives are associated with function. 2.1% of correctly classified non-disease variants are assigned a functional role. Overall, few variants are assigned function, and inclusion of those in a disease classification model would slightly increase the false positive rate.

Investigation of the Role of Protein Structure Accuracy

Two-thirds of the mis-sense mutations are analyzed in the context of structure models rather than experimental structures. The accuracy of these comparative models therefore plays a role in the accuracy of disease assignment. In general, accuracy of a structure model decreases with decreasing sequence identity between the structure of interest and the closest available template structure.

To investigate the significance of this factor, disease assignment accuracy was examined as a function of structure/template sequence identity, in ranges between 25% and 100% ('100%' are those cases for which an experimental structure of the human protein is available). A separate SVM model was trained and tested within each sequence ID group.

Results are shown in Table 2-2. Overall, disease assignment using protein models based on a structure template with more than 40% sequence identity is not significantly less accurate than that based on experimental structures. For sequence identity of 30% or lower, errors in structure models begin to have a significant effect, with increases in both the false negative and false positive rates. Multiple factors contribute to the decline in accuracy, included less reliable side chain interactions arising from higher main chain position errors, an increased frequency of sequence alignment errors, and higher number of insertions and deletions (Tramontano and Morea 2003).

	Disease Variants				Non-Disease Variants			
% Identity	# Mutants	% of Total	# Proteins	FN	# Mutants	% of Total	# Proteins	FP
100%	1710	35%	85	25.5%	672	23%	50	16.7%
90-99%	981	20%	67	23.2%	932	33%	61	13.2%
40-99%	1077	22%	93	24.3%	705	25%	62	16.7%
25-39%	1181	24%	142	***27.5%***	551	19%	91	***28.2%***

Table 2-2. Disease Assignment Accuracy as a Function of Structure Model Quality. Each row shows data using structure models based on a given range of sequence identity to an experimental structure. Accuracy is measured by the false positive rate, FP (fraction of non-disease variants classified as disease causing), and the false negative rate, FN (fraction of disease variants classified as non-disease causing) The '100%' row gives data for cases based on an experimental structure, rather than a model. Accuracy is unaffected by the use of a structure model for sequence identities down to about 40%.

Discussion and Conclusion

Role of Protein Destabilization in Monogenic Disease

This work tested the hypothesis that destabilization of protein structure is a major factor in human monogenic disease. A simple factor based model of the stability impact of single residue mutants and an objective machine learning technique are used. In properly jack-knifed testing, the model is able to distinguish between mutants likely to lead to disease and those that do not, with reasonably low false negative (26%) and false positive (15%) rates. The false negative rate (those non-synonymous base changes that lead to disease not so categorized) partly reflects deficiencies in the model, but also includes the fraction of mutants that act through mechanisms other than destabilization. We conclude from these results that substantially more than half of monogenic disease mutants act through a process consistent with destabilization of the folded state.

Use of the model to classify *in vitro* mutagenesis data strongly supports the role of stability in disease, and implies that a disease causing mutant typically destabilizes a protein by 2 – 3 Kcal/mol. For most globular proteins, the free energy difference between the folded and unfolded state is between 5 and 15 Kcal/mol (Privalov 1979), corresponding to an equilibrium constant between the unfolded and folded states of between 10^{-4} and 10^{-13}. A mutant that destabilized by 2 Kcal/mol would increase the concentration of the unfolded state by about two orders of magnitude, but the fraction of unfolded molecules is still so small that there would be no expected effect on function in an *in vitro* assay. *In vivo,* though, chaperone scavenging of unfolded proteins (Hohfeld et al. 2001) may cause such a 100 fold increase in the fraction of unfolded molecules to result in a much lower steady state protein concentration.

Although loss of stability is clearly highly related to a disease outcome, it may sometimes be an effect on folding which is the actual mechanism. *In vitro* folding studies of simple proteins, such as barnase (Serrano et al. 1992c), show that about 40% of mutants that affect stability also affect the folding rate. For disease mutants, folding may be slowed so much that most molecules are targeted for recycling by the quality control machinery in the ER and elsewhere (Plemper and Wolf 1999). Since not all mutants that affect stability also affect folding rate, if folding were the primary factor, a stability model should generate a high level of false negatives. The reasonably high accuracy of the stability model thus suggests that it is the most relevant factor. Nevertheless, without extensive experimental studies, it is not possible to know for what fraction of cases stability or folding is most relevant.

Direct experimental evidence for the role of stability is scarce, since there are very few studies of the properties of disease-causing mutants in human proteins. One exception is for mutants in phenylalanine hydroxylase. Excess phenylalanine is toxic, and defects in this enzyme lead to phenylketonuria (PKU). Over 100 single residue disease causing mutants are known, and a subset of these have been studied in COS cells (Scriver et al. 2003). There is a clear correlation between the set assigned as affecting stability by our model and the *in vivo* total activity and concentration, as measured by immuno-precipitation.

Why does Protein Stability play a Prominent Role in Monogenic Disease?

There are many mechanisms by which a single base change may affect the function of a protein *in vivo*: Changes in gene regulatory regions may lead to altered transcription rates; changes in the transcribed message may lead to altered processing, particularly splicing; message changes may affect translation through, for example, altering the secondary structure properties (Pelletier and Sonenberg 1987; Shen et al. 1999). Surprisingly, data for monogenic disease in HGMD suggest that all these pre-protein factors account for less than 10% of cases (Stenson et al. 2003). This number may be an underestimate of the true value, because of bias in detection methodology. Nevertheless, it is clear that protein level effects are by far the more common.

Once a polypeptide chain has been produced, non-synonymous changes (those base changes resulting in an amino acid substitution) may affect *in vivo* activity in two major ways: Aspects of the protein's molecular function may be altered, particularly ligand binding, catalysis, post-translational modification, or an allosteric mechanism. The likelihood of this class of effect depends on the fraction of residues in critically involved in one or more of these functions.

The second way in which non-synonymous base changes may affect *in* vivo activity is by reduction of the concentration of protein. This may come about through less successful folding, or an increase in the fraction of unfolded protein, caused by a reduction in stability. Tests with the stability model, sampling a large number of randomly chosen mutants, show that approximately half are consistent with a disease outcome. Thus, the high fraction of disease mutants associated with stability loss is likely a consequence of the higher fraction of mutants that can affect stability, compared with the other possible causes.

Distinguishing Properties of Monogenic Disease Proteins

For the 1000 or so monogenic disease proteins in HGMD, the average number of known single residue mutants leading to disease is just over 10 (Stenson et al. 2003). Yet no mutants directly causative of monogenic disease are known in the remaining approximately 22,000 human proteins. What is the difference between these two sets of proteins? First, monogenic disease proteins may be abnormally unstable or have abnormally fragile folding behavior. There is very little data with which to address this possibility, but many are relatively simple metabolic enzymes, and compared with most human proteins, the least likely to exhibit this sort of fragility. A second possibility is that mutants in many of the other proteins lead to a non-viable fetus, and so never be classified as disease causing. Gene suppression in C.elegans (Kamath et al. 2003) and Saccharomyces (Cliften et al. 2003; Rubin et al. 2000), as well as limited mouse knockout data all suggest that only 10 - 20% of proteins are essential in this sense, and so that is unlikely explanation. Third, and most probably, monogenic disease proteins may be the subset to which the system is least robust to component failure. Analysis of non-synonymous single nucleotide polymorphisms in the human population shows a significant fraction that appear to be as deleterious to protein structure and function as those found in monogenic disease genes (Chasman and Adams 2001; Ng and Henikoff 2003; Ramensky et al. 2002; Yue 2005), but with no disease outcome. Limited knowledge of human protein networks makes it difficult to rigorously test this possibility. Nevertheless, inspection of the pathway context of monogenic disease proteins supports this explanation. Many, such as phenylalanine hydroxylase, appear to perform unique roles, with no redundancy alternative pathways. In contrast, inspection of the pathway context of proteins that contains SNPs that destabilize protein structure significantly, such as the T cell receptors (Wang and Moult 2003), usually suggests a mechanism that makes the system robust to failure of a protein component. Many different T cell receptors are involved in an antigenic response, so that reduced effectiveness of some will not have obvious disease consequences, although it may influence resistance to particular infections in subtle but significant ways.

Advantages and Disadvantages of a Protein Structure Based Approach

An advantage of the structure based approach is that it provides a detailed atomic level model of the precise mechanism by which an amino acid change results in a change in protein properties. A disadvantage is that it is limited to stability effects. Use of comparative modeling allowed us to extend the number of mutants that can be analyzed. Tests showed that disease prediction accuracy is unaffected by the use of a model, down to 40% sequence identity to a known structure. This is in keeping with studies of the accuracy of structure modeling methods (Cozzetto and Tramontano 2005), and also partly reflects the fact that the method does not depend on very accurate structures. Even so, only about 10% of human protein domains can currently be analyzed. The rapid advance of structural genomics (Service 2005) may quickly reduce this limitation.

Methods

Identification of Single Residue Variants related to Monogenic Disease

Genes associated with monogenic disease were identified by checking all 16,220 human gene names in the NCBI Locuslink (Wheeler et al. 2004) database (as of 04/26/2002) against the Human Gene Mutation Database(Stenson et al. 2003), (HGMD) (as of 02/09/2002). HGMD contains the most comprehensive collection of mutations related to monogenic disease. Most are causative of monogenic disease, although a few may be associated with disease as a result of linkage disequilibrium rather than directly causative, or contribute to complex trait disease. Later versions of HGMD include more of the latter class, and so the earlier version was preferred. 731 genes containing 10,263 single residue variations were identified.

Identification of a Set of Single Residue Variants not related to Disease

We also required a control set of mutants, not causative of disease. It is not known which base variants in the human population contribute to complex trait disease, and so it is not possible to use these. Following others (Sunyaev et al. 2001), we used non-synonymous base differences between human proteins and closely related proteins in other mammals. The justification here is that almost all variants that are fixed between species are essentially neutral and non-deleterious. To maintain compatibility between the disease and control sets, the same 731 monogenic disease proteins were used. The protein sequences of these genes were compared to all other mammalian protein sequences in Swiss-Prot (Boeckmann et al. 2003), using BLAST (Altschul et al. 1997). Proteins with at least 90% sequence identity over at least 80%

of the full length were selected. Single residue differences in these alignments were used as a set of pseudo ‘mutations’, providing the non-disease set. A total of 348 proteins containing 16,682 such single-residue differences to the human disease set were obtained.

Selection of Sets of Mutants with Protein Structure

Each of the 731 human proteins was checked for entries in the Protein Data Bank (as of 7/26/2004)(Deshpande et al. 2005). Templates for models of human proteins were taken from the PDB for cases where there was no human structure available, and there was a PDB entry for an X-ray structure at least 3.0 Å resolution and with 40% or higher sequence identity to the human protein over at least 100 residues.

For the non-disease set, variants that might be partially compensated by other species differences in the same protein were eliminated as follows. All clusters of variants where there are interatomic contacts of 5Å or less between residues were discarded. For example, A2S in the myosin light chain is a variant between human and mouse, and between human and rat. G20T, a variant between human and mouse, makes contact with the G20 position, and so both variants were excluded. The rat protein has no change at G20, so rat A2S was retained in the non-disease set.

Support Vector Machine

The Support Vector Machine software package SVMlight (http://svmlight.joachims.org/) was used to determine the partitioning surface between the disease and non-disease data in the 15 dimensional parameter space.

Continuous variables were normalized in the form of a Z score [Z=(value-mean)/standard-deviation]. A radial basis kernel was used, allowing for complex surface topology. For this kernel, the higher the parameter γ, the more complex the effective surface, allowing better accommodation of the data. Too higher a gamma leads to over-fitting, and less accurate prediction on new data. A γ value of 0.2 was selected, based on a series of trials. Weights were assigned to the disease and control data sets to compensate for their different sizes, such that they contributed equally to determining the partitioning surface. The distance of a data point from the partitioning surface provides an approximate measure of confidence in a prediction.

SVM Model Training and Testing

A subset of 90% of the disease and non-deleterious variations were selected randomly to form a training set. The resulting SVM model was used to predict which of the 10% of data not included in training are disease causing. The training and testing procedure was repeated 30 times, randomly selecting the test data on each run. For each trial, the false negative rate (the fraction of disease variations mis-classified as non-disease) and false positive rate (the fraction of non-disease variants mis-classified as disease causing) in the test dataset were calculated. The average false positive and false negative rates provide the measure of the prediction accuracy.

In-vitro Mutagenesis Data

Free energy difference data from site directed mutagenesis experiments was used to test the ability of the SVM model to identify known destabilizing mutations. Four proteins with a large number of associated site directed mutagenesis experiments(Byrne et al. 1995; Fulton et al. 1999; Green et al. 1992; Green and Shortle 1993; Holder et al. 2001; Kragelund et al. 1999; Main et al. 1998; Meeker et al. 1996; Schwehm et al. 1998; Serrano et al. 1992a; Serrano et al. 1992b; Serrano et al. 1992d; Shortle et al. 1990; Stites et al. 1994) were selected. They cover three classes of protein folds (SCOP (Andreeva et al. 2004) classification): all alpha, all beta and alpha & beta. Table 2-1 lists the proteins, and the number of mutants in each. More data are available in the PROTHERM database (Bava et al. 2004) but inconsistencies in format, particularly the sign convention for free energy, prevent the large scale use of these.

Comparative Modeling of Protein Structure

Comparative models were built using the in-house APSE (Automatic Protein Structure Emulator) pipeline. Modeling protocols in APSE are based on experience with building comparative models in the CASP experiments (Samudrala and Moult 1997) and a variety of projects. The procedure can be run in automatic or semi-automatic mode. A core backbone model is first constructed by copying regions of the chosen template structure. Alignments are obtained using CLUSTALW. Co-ordinates of side chains conserved between the human protein and the PDB template are copied. Remaining side chains are added using SCWRL (Canutescu et al. 2003).

Where necessary, quaternary structure was taken from the PQS (Protein Quaternary Structure) database of biological units (Henrick and Thornton 1998). Additional subunits are modeled in the same manner as the chain of interest. Side chains are modeled in the multimer context.

Modeling the Structure of Single Residue Mutants

All SCWRL library back-bone dependent conformations (Canutescu et al. 2003) for the new side chains were built. The conformation least damaging to stability was selected, based on the following rules. First, the conformation with the least worst overpacking was selected. I.e. if there is one conformation with an interatomic contact of 2.6Å and another with 2.7Å, the latter was accepted. No distinction was made between conformations with contacts 3.0Å or longer. If more than one conformation remained, the one with the least loss of hydrophobic area was selected. In cases where there is no loss of hydrophobic area, conformations with loss of a salt bridge were next eliminated, then those with electrostatic repulsion, hydrogen bond loss, cavity formation, backbone strain, introduction of a buried charge, and finally, introduction of a buried polar group.

Modeling the Stability Impact of a Single Residue Mutant

Table 2-3 lists the stability factors that provide the 15 dimensions used in assessing the impact of each mutant on protein stability. These are divided into those factors treated as continuous variables, and those treated as two state variables (significantly destabilizing or not).

Continuous factors

1. Electrostatic interactions: The difference in electrostatic energy between a wild type protein and its corresponding mutant was calculated using a simple Coulomb's law treatment, with no solvent model. The partial electrostatic interaction energy between a pair of polar or charged groups 'i' and 'j' is calculated in the usual manner as:

Type	**Factors**
Continuous factors	Electrostatic interaction: polar-polar, polar-charge, charge-charge.
	Over-packing
	Hydrophobic burial
	Surface accessibility
	Structural rigidity: Crystallographic B-factor, Z score and standard deviation
Binary factors	Cavity
	Electrostatic repulsion
	Backbone Strain
	Buried charge
	Buried Polar
	Breakage of a disulfide bond

Table 2-3. The 15 factors included in the stability model.

The effect of each single residue mutant on stability is expressed in terms the value of one or more of these contributions to the energy and entropy. ‘Continuous factors’ are represented by a continuous variable, ‘binary factors’ are two states, either significantly or not significantly affecting stability

$$E_{ij} = K \sum_k \sum_l q_k q_l / r_{lk}$$

where the sums are over all atoms 'k' of group ‘i’ and atoms 'l' of group ‘j’, the 'q's are the partial atomic charges in electrons, and r_{lk} is the distance between atoms 'l' and 'k', in Å. 'K' is the scaling constant (332) nominally converting energies to Kcal/mol. (Absolute scale is not significant here, because of the Z score normalization). Interactions between a pair of groups are included if the centers of charge are less than a cutoff distance d_c apart. The center of charge of a group $\mathbf{r}_c$ is defined as:

$$\mathbf{r_c} = \sum_k |q_k| \mathbf{r}_k / \sum_k |q_k|$$

Where the sum is over all atoms in the group. Electrostatic group definitions and partial atomic charges are as in (Pedersen and Moult 1997). The threshold for group-group interactions, $d_{c,}$ is 5Å. This protocol for electrostatic calculations has been to be effective at identifying incorrect structural features in experimental structures(Oliva and Moult 1999).

2. Overpacking: For each mutant, the closest inter-atomic distance between the mutant residue and any neighboring residue was used.

3. Relative surface accessibility: Solvent accessible surface (Lee and Richards 1971) was calculated with in-house software. The relative surface accessibility of a residue is defined as the surface area of the side chain in the folded state divided by an estimate of the average surface area in the unfolded state (Shrake and Rupley 1973).

4. Hydrophobic burial Change: The change in buried non-polar area ΔA_{NP} resulting from a single residue mutation is defined as:

$$\Delta A_{NP} = \sum_i \Delta a_i - \sum_j \Delta a_j$$

where the first sum 'i' is the change in non-polar area on folding for all non-polar atoms in the mutant structure and the second sum 'j' is over all non-polar atoms in the wild type structure. The change in atomic non-polar area in folding is given by:

$$\Delta a = a_u - a_f$$

where a_u is the estimate of the average atomic surface area in the unfolded state (Shrake and Rupley 1973) for that atom, and a_f is the calculated atomic area in the folded structure. Non-polar atoms are those assigned zero charge.

5. Crystallographic Temperature factors: For each experimental structure used directly or as a model, the average temperature factor <B>, and standard deviation σ(B) over all Cα atoms was calculated, and used to obtain a temperature factor Z score for each Cα: $Z_i = (B_i - \langle B \rangle)/\sigma(B)$. B_i, Z_i, and σ(B) were used as parameters in the SVM.

Binary factors

6. A Cavity is assigned to any mutation resulting in the loss of volume of an aliphatic carbon group or greater at a zero solvent accessibility position. For example, Ala mutated to Gly, where the wild type Cβ atom has zero solvent accessibility.

7. Electrostatic repulsion is assigned to any mutation which results in two like charged groups with an unavoidable atomic contact of less than 4.5 Å.

8. Backbone Strain is assigned to any mutation if one of the following conditions is met:

A. Replacement of a glycine residue with φ/ψ angles in a non-allowed region for other residue types. Allowed regions were those covering 90% of observed φ/ψ values, as provided in PROCHECK (Laskowski RA 1993),

B. Replacement of a cis-proline (ω = 0+/-60^o) with another residue.

C. Replacement of another residue by proline, where the φ value is inappropriate (Permitted φ for Pro = -60+/-15^o).

9. Buried charge is assigned to any mutation which results in a zero solvent accessibility, electrostatically isolated, charge group.

10. Buried Polar is assigned to any mutation which results in a zero solvent accessibility polar group with no hydrogen-bond. A hydrogen bond is defined as a donor to acceptor distance <= 2.5Å, and an angle at the acceptor >= 90.0º.

11. Breakage of a disulfide bond is assigned to any mutation which replaces a cysteine residue in an S-S bond with a non-cysteine residue.

Evaluation of Discrimination power of each Stability Factor

The frequencies of each stability factor in the disease and non-disease datasets were calculated. The ratio of the two frequencies defines a discrimination power. For this purpose, a threshold was chosen for each of the continuous factors. Any mutation with a value higher than the threshold was considered to destabilize protein structure. Thresholds were chosen by inspection of the distribution of values for the disease and non-disease sets, selecting levels that provide a high fraction of true positives and true negatives, while minimizing false negatives and false positives. The following values were used:

1. Overpacking: At least one unavoidable atomic contact of 2.5Å or less of the mutated residue to a neighboring one.

2. Hydrophobic burial: Loss of hydrophobic burial of more than 50 $Å^{2.}$

3. Electrostatic interaction: Any reduction in electrostatic interaction energy, for polar-polar, charge-charge and charge-polar interactions.

4. Buried residue: Relative residue accessibility of less than 20%(i.e. the wild type side chain accessibility is less than 20% of the estimated average unfolded state accessibility).

5. Moderate crystallographic Temperature Factor: The Cα temperature factor of the mutated residue has a Z score of less than +1 (i.e. the temperature factor is less than one standard deviation above the mean for the protein).

Identification of Residues with a Role in Molecular Function

Each mutated residue, and all residues with one or more atomic contacts of 6Å or less to it, was checked against the SWISS-PROT feature annotation table for possible functional effects. Additionally, a check was made for atomic contacts of the mutated residue of 6Å or less to any ligand atom in PDB entries for that protein and other X-ray structures with at least 40% sequence identity over at least 100 amino acids, and at 3.0Å or better resolution.

Chapter 3: Identification and Analysis of Deleterious Human SNPs

Introduction

In this chapter, we analyze non-synonymous or missense SNPs in the human population. We use two methods to identify which missense SNPs are deleterious to protein function. Both methods have been developed and tested on amino acid changes causative of monogenic disease, and a control set of single residue changes fixed between closely related mammalian species (Sunyaev et al. 2001). One method analyzes the impact of amino acid changes on protein stability, making use of the three dimensional structural environment (Yue et al. 2005) as described in the chapter 2. We find the majority of single base changes that cause monogenic disease significantly destabilize the folded state of the protein concerned. The second method, reported in this chapter, makes use of the tendency of critical amino acids to be conserved within a protein family. The more conserved and restricted the type of amino acid at a position, the more likely that a substitution not consistent with that pattern will have a deleterious impact on protein function. This method is more general than the stability model, including all types of protein level effect. It is also more widely applicable, since it does not require knowledge of three dimensional structure. On the other hand, it provides less direct insight into the mechanism by which a missense SNP affects protein function. The principles of sequence conservation methods have also been explored by others (Chasman and Adams 2001; Krishnan and Westhead 2003; Ng and Henikoff 2003; Ramensky et al. 2002; Sunyaev et al. 2001). We have used a machine learning method, the support vector machine, trained on five simple features that capture the relative sequence conservation at each position in a multiple sequence alignment. The support vector machine allows the identification of a subset of high confidence predictions. The use of two separate methods provides an additional means of assessing the reliability of the conclusions.

The two methods have been used to analyze sets of non-synonymous SNPs found in the human population, extracted from the dbSNP database (Sherry et al. 2001), and a subset of those for which population frequency data are available. The subset are data from Perlegen (Hinds et al. 2005) and the Hapmap consortium (2003). Using stringent criteria, we find that about ¼ of these SNPs are classified as deleterious at the same level as those causing monogeneic disease in other genes. These are very likely to have a significant impact on protein function, and so probably contribute to complex disease traits, and provide a basis for prioritization in association studies.

We have also examined a number of aspects of the relationship between monogenic disease genes and the rest. First, we have compared the occurrence of deleterious SNPs in monogenic versus non-monogenic disease genes. We find that, whereas in monogenic disease genes nearly all deleterious SNPs occur at low frequency in the population, in other genes, a larger proportion are found at high frequencies, consistent with the idea that the effect of deleterious SNPs in other genes is buffered. Second, we have looked at the rate of sequence divergence of monogenic versus other genes. An interesting variation with conservation level is found. Third, we have found that there is a correlation between the phenotypic impact of mouse knockouts and whether or not the orthologous human gene is implicated in monogenic disease. Finally, we have checked to see if monogenic disease genes are less likely to have paralogs than the others, exploring the idea that paralogs some times can provide substitute function. No such effect is found.

Results

Training and Testing Data used for the Classification Methods

Table 3-1 summarizes the monogenic disease and control datasets used for training and benchmarking the sequence profile and structure stability methods. There were a total of 10263 deleterious mutants in 731 proteins and 16682 control substitutions in 348 proteins available. The profile model includes 92% and 71% of these respectively, since profiles can be built for most proteins. In testing, high confidence ('HC', SVM score > |0.5|) classifications were obtained for over 80% of these. Significantly fewer data (37% and 14% respectively) are included in the stability model, because of low structural coverage of human proteins. High confidence classifications are again obtained for about 80% of cases. The last two rows show the data for cases where both methods could be applied. The fraction of high confidence predictions is similar.

Accuracy of the Classification Methods

Figure 3-1 shows the false positive (blue bars) and false negative rate (red bars) for both methods separately, on all data and for just the high confidence classifications (an SVM score of greater than 0.5 for non-deleterious classifications and less than -0.5 for deleterious ones), as well as the corresponding data for the cases where the two methods agree. As expected, the false positive and false negative rates are highest for the individual classification methods, lower when only high confidence classifications are considered, and lowest of all when only high confidence classifications shared by both methods are included (3% false positive,

	Deleterious Mutants			**Control substitutions**		
	#	**%**	**Proteins**	**#**	**%**	**Proteins**
All Data	10263	100%	731	16682	100%	348
Profile	9468	92%	693	11778	71%	336
Profile HC	7986	78%	673	10171	61%	336
Stability	3768	37%	243	2309	14%	153
Stability HC	3046	30%	229	1904	11%	152
Profile+Stability	3641	35%	235	2141	13%	148
Profile+Stability HC	2501	24%	216	1498	9%	146

Table 3-1. Training and Testing Data for the Profile and Stability methods. ‘Deleterious mutants’ are amino acid changes that cause monogenic disease (Stenson et al. 2003). ‘Control substitutions’ are amino differences between human proteins and closely related orthologs. ‘HC’ are high confidence classifications from the Support Vector Machine. The ‘#’ and ‘%’ columns give the number and percent of data in each row. ‘Proteins’ are the number of proteins from which data are included.

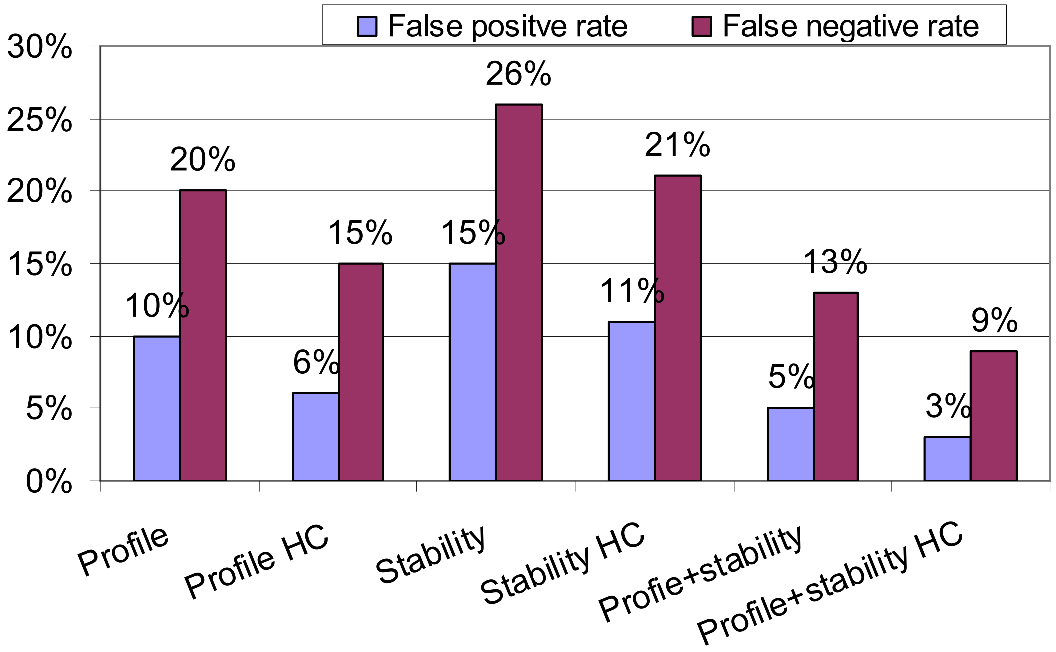


Figure 3-1. Evaluation of the profile and Stability Methods.

False positive and false negative rates are shown for the two methods alone, and for cases where both can be applied and the classifications agree. Results are shown for all classifications, and for the high confidence subsets (‘HC’, SVN score > |0.5|). Higher false negative rates for the stability model reflect the fact that only stability and folding effects are included, where as the profile model includes all effects on protein function *in vivo*.

9% false negative). The false negative rate of the profile method is slightly lower than that of the stability method (20% versus 26% for all classifications, 15% versus 21% for high confidence ones). This difference is expected, since the profile method includes all effects on protein function at the amino acid level, including ligand binding, catalysis, allosteric mechanisms, and post-translational modifications, as well stability and folding effects, whereas the stability model includes only stability and folding contributions. Less expected is the lower false positive rate for the profile method (10% versus 15%, 6% versus 11% for high confidence classifications). The balance between false positive and false negative rates is determined by the relative weights given to the deleterious and control datasets in training the SVM. Equal weights, taking into account the differences in data set sizes, were used. Different weighting would adjust the false positive rates to be more similar.

For both methods, the finite error rates reflect both the effects of approximations in the methods and the nature of the data sets. The stability method incorporates a number of approximations in modeling the structure of mutants, and uses a scenario based analysis of effects on stability (Yue et al. 2005). As discussed later, for the profile method, the effect of a limited number of sequences in a profile is the main approximation. The HGMD data (Stenson et al. 2003) used as a disease set contains some entries that are not strictly causative of monogenic disease. For example, the mutant G15D in the alpha chain of Hemoglobin (HBA1) is in HGMD, but is predicted to be non-disease causing, with a confident SVM score of 2.9. The literature on this mutation (Molchanova et al. 1994) gives no indication of disease. Since 1999, HGMD have added some mutants that disease 'associated' or 'risk' polymorphisms. This work uses the HGMD version of 04/26/2002, which includes 152 mutants identified as not necessarily causative of disease. The false negative rate for these is very high: 62% for the profile method and 73% for the stability method. The assumption of no deleterious effects fixed between species might contribute to a finite false positive rate. There are 41 HGMD mutants where the altered amino acid is the wild type in another species. Of these 37 are classified by the profile model, but only four are found to be deleterious. Thus, these appear to be largely inappropriate entries in HGMD, rather than deleterious mutants fixed in other species. In spite of the limitations in the models and data, the errors are sufficiently small that firm conclusions about the level of deleterious SNPs in the human population can be reached, provided the false positive and false negative rates are taken into account.

Sensitivity to the Number of Sequences in a Profile

To classify deleterious nsSNPs, the profile method makes use of five features related to the relative sequence conservation, including the probability of accepting an amino acid substitution (based on PSSM, position specific scoring matrix) and four entropy factors. The reliability of the PSSM and entropy values depends on the size of the sequence alignment. We examined the accuracy of the method as a function of the number of sequences available, after filtering out redundant and less reliably aligned sequences, as described in Methods. Profiles were divided into sets with different numbers of sequences, and the accuracy evaluated for each set. Table 3-2 shows the results. All sets have similar accuracy, except the set with the smallest number of sequences (2 – 9). This group has a similar false negative rate but a higher

No. of sequences	Deleterious			Control		
	FN	Proteins	#	FP	Proteins	#
[2-9]	0.18	60	500	0.31	16	787
[10-19]	0.17	82	1073	0.14	24	1296
[20-39]	0.18	167	1957	0.11	85	2785
[40-59]	0.22	121	1871	0.13	66	2263
[60-79]	0.19	94	804	0.10	48	1578
>=80	0.18	169	3263	0.10	97	3069

Table 3-2. Accuracy of the Profile method as a function of the number of sequences in the alignment.

Accuracy is measured in terms of the false negative rate (FN) and the false Positive rate (FP). The '#' columns show the number of variants analyzed in each alignment size range, and 'proteins' are the number of human proteins included. Accuracy is approximately equal in all but the smallest alignment range, where there is a sharp rise in the false positive rate.

false positive rate (31%) than the other groups. The high false positive rate is probably a consequence of the low maximum entropy for a small number of sequences: the maximum for two sequences is approximately 1 bit, while for 20 sequences, it is 4.3 bits. Additionally, for small profiles, the PSSM is dominated by the BLOSUM scores rather than the pattern of residue use.

Comparison between BLOSUM, PSSM, and Profile Models

The full profile method includes the PSSM for the aligned sequences, and entropy factors. We compared the performance of a PSSM (Altschul et al. 1997) alone, which takes into account which residues are observed at each position in a sequence alignment, with performance using an average substitution matrix (BLOSUM (Henikoff and Henikoff 1993)), which considers only the likelihood of the substitution in all proteins at all positions. It has been suggested that the BLOSUM matrix is suitable for use in identifying damaging nsSNPs (Cargill et al. 1999; Ferrer-Costa et al. 2002). Since a PSSM contains information unique to each sequence family and sequence position, we would expect it to produce more accurate classifications.

BLOSUM 45 and BLOSUM 62, representing average substitution preferences between proteins with different levels of sequence identity, were tested. PSSM and BLOSUM method accuracy as a function of a score threshold were examined, and the threshold returning the lowest sum of false positives and false negatives chosen in each case. The results are shown in table 3-3, with the full profile method included for comparison. The BLOSUM matrices both yield similar false positive and negative rates, of about 27% and 36% respectively, whereas the PSSM has significantly lower

	False Positive Rate	**False Negative Rate**
BSOSUM 45	27%	38%
BLOSUM 62	28%	36%
PSSM	22%	28%
Profile model	10%	20%

Table 3-3. Comparison of Classification accuracy of BlOSUM matrices, a PSSM and the full Profile method.

The PSSM method has substantially lower false positive and false negative rates than obtained with either BLOSUM matrix. The additional entropy information in the full profile model further improves accuracy.

values of 22% 28%. The profile model is substantially more accurate than the PSSM alone, with false positive and false negative rates of 10% and 20% respectively, establishing the entropy terms do provide significant additional information.

Comparison of Expected and Observed Accuracy for the Combined Classification Methods

Reliable false positive and false negative rates are essential for accurately estimating the fraction of deleterious SNPs in the population. In principle, the values obtained from the benchmarks are accurate. A further test is provided by examining the consistency of the individual method errors with those of the combined methods. Since the two methods are based on different principles and their primary causes of error are unrelated, the errors are approximately independent. Under these conditions, the expected specificity and sensitivity for the cases where the methods agree can be calculated (see Methods). Comparison of these values with the actual ones then provides the consistency test.

Table 3-4 shows the observed and expected values for the combined methods. Results for two combined sets are shown. The first includes the subset of data to which both methods can be applied. Specificity is substantially higher for the combined methods, and sensitivity lower, as expected. The second set includes data to which both methods can be applied, and where high confidence classifications are returned in all case. For both data sets, the expected and observed values are reasonably close, supporting the reliability of the false positive and false negative rates returned by single method benchmarking. Observed sensitivities are a little higher than expected, suggesting the false negative rate may be slightly

		Single Model		Both Models	
		Profile	**Stability**	**Observed**	**Expected**
All Results	**SN**	80%	74%	63%	59%
	SP	90%	85%	96%	98%
High Confidence	**SN**	85%	79%	73%	67%
	SP	94%	89%	97%	99%

Table 3-4. Sensitivity (SN) and Specificity (SP) for the combined methods, compared with that Expected from the accuracy of the individual ones.

The first pair of columns shows the sensitivity and specificity for each method alone. The third column shows the results for the cases where both models can be applied, and the fourth column shows the expected sensitivity and specificity of the combined models, given the results for the individual methods. The first two rows show the results for all classifications. The high confidence set includes only cases where a high confidence result is obtained. Agreement between observed and expected provides a test of the accuracy values (see Methods).

overestimated. Specificities are about 2% lower than an expected, but the values are so high (worst case 96%) that any small amount of noise may account for this.

Analysis of Population SNPs: Approximately a Quarter of Non-Synonymous Population SNPs are Deleterious.

We now use the profile and stability methods to identify deleterious non-synonymous SNPs (nsSNPs) in the human population. As described in Methods, nsSNP data were obtained from three sources: the NCBI dbSNP database (Sherry et al. 2001), the Perlegen data (Hinds et al. 2005), and the Hapmap project results (2003). dbSNP contains a wide range of data, some of which is based on a single observation. Both Perlegen and the Hapmap project have genotyped sets of individuals from several different populations. Since these SNPs are all verified, and have associated population frequency information, we have analyzed them as a separate data set, referred to as the Frequency set. Table 3-5 shows the number of data available in the full dbSNP set and the Frequency set, and the number of data that can be classified by the stability and profile methods, the combined methods, and the number of high confidence classifications in each case.

Figure 3-2 shows the fraction of population SNPs assigned as deleterious in dbSNP (blue bars) and the Perlegen/Hapmap data (purple bars). Results are again shown for the two methods separately, and for the combined methods, for all classifications, and those of high confidence. Deleterious classifications in both SNP sets are lowest for the most stringent conditions (high confidence classifications for the combined methods), with 33% for all dbSNP data and 17% for the Frequency subset, The highest deleterious rates are for the stability model alone, with 40% for

	All			Frequency Set		
	# SNPs	% SNPs	# Genes	# SNPs	% SNPs	# Genes
dbSNP Build 124	50772	100%	15710	10403	100%	6316
Profile	29081	57%	11129	6377	61%	4297
Profile HC	22067	43%	9782	4911	47%	3549
Stability	5166	10%	2019	885	9%	624
Stability HC	3960	8%	1776	681	7%	509
Profie+stability	3150	6%	1512	531	5%	415
Profile+stability HC	2096	4%	1180	370	4%	304

Table 3-5. Data used for identifying deleterious Human SNPs.

The top line shows the number of missense SNPs available in the dbSNP database, and the subset of these with population frequency information, from Perlegen and the Hapman project. Classifications were made on the full set and the frequency set. The number of SNPs classified in each case, and the number of genes are given for the profile method, the stability method and the combined methods. In each case, values are given the full data and for the subset that are classified with high confidence (SVM score > |0.5|).

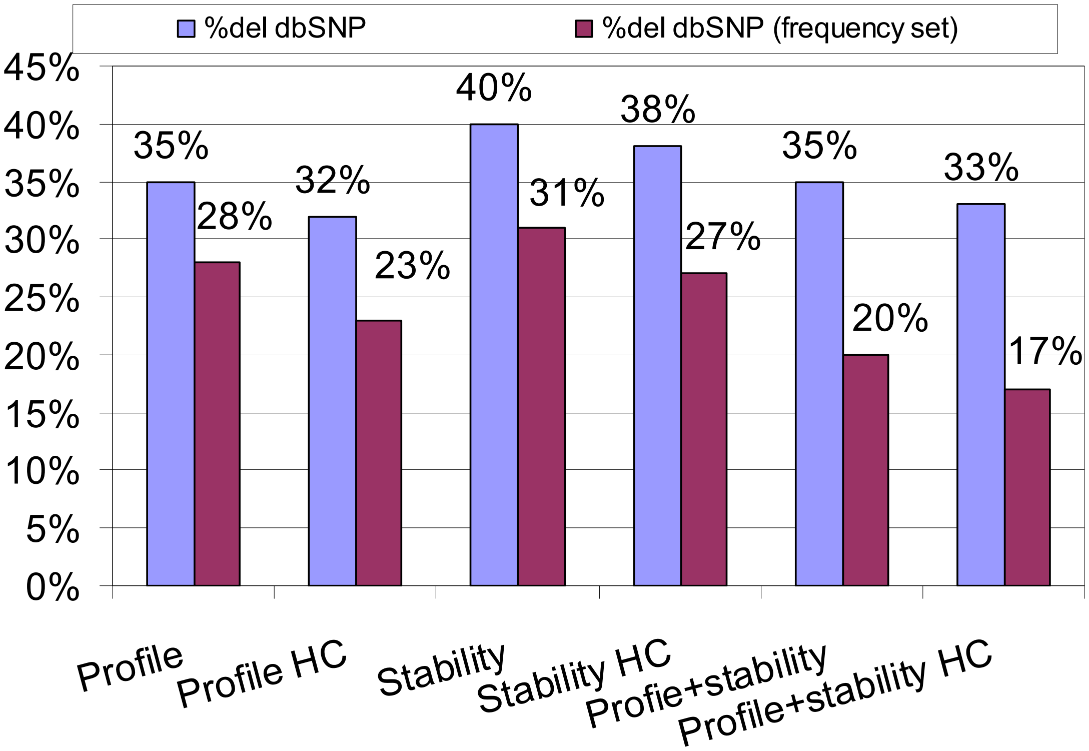


Figure 3-2. Estimated fraction of Deleterious SNPs in the Human Population. Results are shown for all missense SNPs in dbSNP build 124 (blue bars), and a subset for which there are population frequency data (purple bars). Deleterious rates are calculated using the profile and stability methods, the two methods combined, and also, in each case, for high confidence ('HC', SVM score > |0.5|) classifications only. Consistently lower rates are found for the frequency subset than for all dbSNP data, partly reflecting the effect of incorrect entries in the latter. Variations in the rate for the different classification methods reflect the differing false positive and false negative levels. Lower rates for the high confidence predictions reflect the fact these are generally obtained only for more severe effects on protein structure and function.

the dbSNP data, and 31% for the Frequency subset. Deleterious SNP rates are consistently substantially lower for the Frequency subset than the full dbSNP set, presumably reflecting the effect of the unreliable single observation component in dbSNP. As a control, we also analyzed the 952 Hapmap SNPs which were found to have zero frequency, that is, are in dbSNP, but were not observed in the Hapmap population. The profile method classifies 50% of those SNPs as deleterious, a much higher value, and close to that obtained in tests introducing random mutations.

The deleterious population SNP rates in figure 3-2 are distorted somewhat by the finite false positive and false negative rates of the classification methods. Distortions can occur in both directions: A high false positive rate contributes to over-estimating the deleterious SNP level, but a high false negative rate contributes to an underestimate. We correct for these effects as follows: For a given true deleterious rate D_{true}, with a false positive rate f_p and false negative rate f_n, the expected apparent deleterious rate D_{exp} is given by:

$$D_{exp} = D_{true} - D_{true} * f_n + [1- D_{true}]* f_p$$

Where the second term (D_{true} * f_n) is the underestimate effect of false negatives, the third ([1- D_{true}]* f_p) is the over-estimate effect of false positives. The most probable value of D_{true} can then be obtained by examining the difference between the expected (D_{exp}) and observed (D_{obs}) deleterious rates, as a function of D_{true}.

Figure 3-3 shows the residual $|D_{obs} - D_{exp}|$ as a function of possible D_{true} values, for each of the method conditions, using the frequency subset. There are well defined minima in the residual curves, at values of D_{true} between 15 and 25%. Lower values are obtained with the high confidence subsets (~20%), and the lowest value (15%) is obtained with the high confidence assignments common to both methods. It is expected that high confidence scores are only obtained for the more severe effects on protein function and stability. Application of the stability method to site directed mutagenesis data where experimental folding free energies are available confirms that on average high confidence assignments have a more severe effect in protein stability (data not shown), Thus, the lower level (15 – 20%) of deleterious SNPs found for the high confidence score subsets are an estimate of the fraction of more severely deleterious SNPs in the population. The best estimate of the fraction of population missense SNPs that are as detrimental to protein function as those found for monogenic disease is provided by the full set of classifications for the profile and stability methods, separately and combined. In all three cases, that value is close to 25%. Thus, the analysis leads to the conclusion that approximately one quarter of non-synonymous SNPs found in the population are as deleterious to protein function as single base changes known to cause monogenic disease. This value is little lower than reports by other groups (Chasman and Adams 2001; Ramensky et al. 2002), probably because of the effect of correcting for finite errors rates in the methods.

Deleterious SNPs in Monogenic Disease Genes

There are 4,458 nsSNPs in dbSNP located in monogenic disease genes, among which 1,656 are assigned as deleterious by the profile method. Only a small

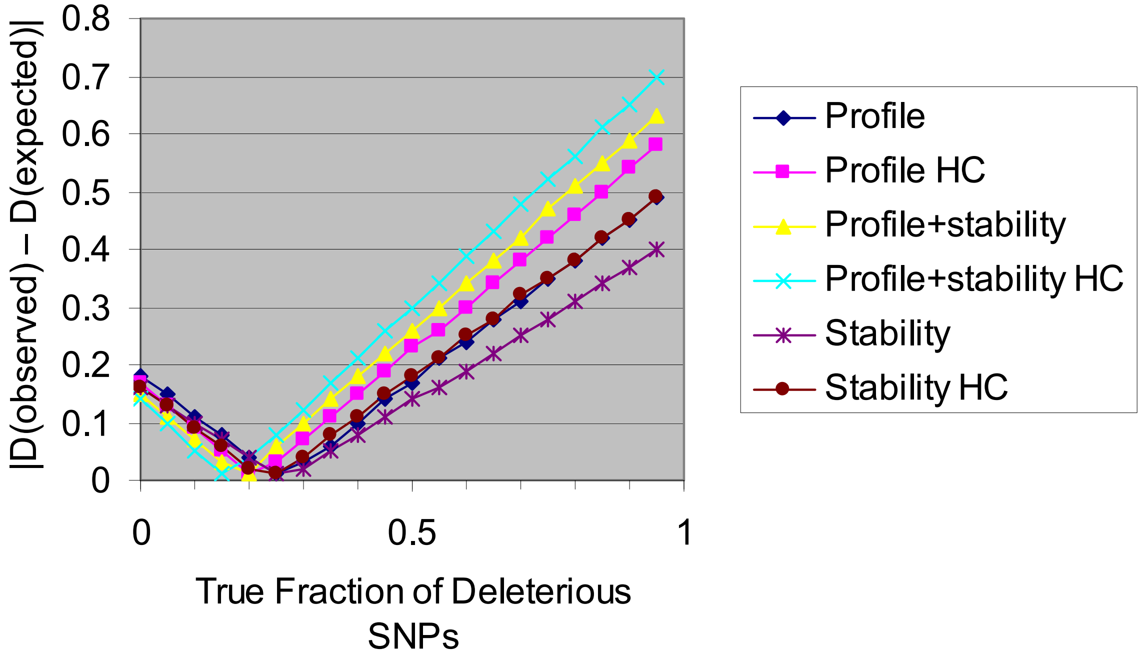


Figure 3-3. Difference between the expected and observed fraction of deleterious population SNPs as a function of the underlying true rate.

Residuals are calculated as a function of possible true values (X axis), using the false positive and false negatives rates for each method. Minima give the estimated true deleterious rates. The Stability, Profile and combined methods all yield rates close to 25%. The High confidence classifications yield lower values, reflecting the fact that generally only severe effects on protein structure and function have high confidence classifications. Data are for the frequency subset of dbNSP.

portion (152) is also present in HGMD as known monogenic disease mutants. The reminder might be new monogenic disease causing variants, known variants not yet entered into HGMD, or false positives. Given a false positive rate of 10%, we only expect 446 in that category. If the additional SNPs really are disease causing, we would expect them to be predominantly at low frequencies in the population. Figure 3-4 shows a comparison of the population frequency distribution of the 970 of these in the frequency subset with the corresponding distribution for all other genes. As expected, there are many more low frequency SNPs in both sets. Both sets also show a higher fraction of deleterious SNPs at low frequency, compared to non-deleterious, consistent with their being selected against. That bias is stronger for the monogenic disease gene set, and only about 10% are at frequencies higher than 20%, the expected fraction of false positives.

To investigate the possibility that some of the additional deleterious missense SNPs in monogenic disease genes are in fact disease causing, we examined the subset of 18, in 15 different genes, which are assigned as deleterious with high confidence by both classification methods. Table 3-6 summarizes the data for these SNPs. Five are already in HGMD, but given the very low false positive rate for this subset (3%, as shown in figure 3-1), the others are candidate mutants for monogenic disease. Two of these have surprisingly high population frequencies for monogenic disease mutants: SEROINA7 L303F, at 20%; and AMACR G175D with a frequency 34%. SERPINA7 belongs to a family of serine protease inhibitors, but also functions as a thyroid binding–globulin (TBG). There are many mutations associated with TBG deficiency, and many of these also have a high population frequency

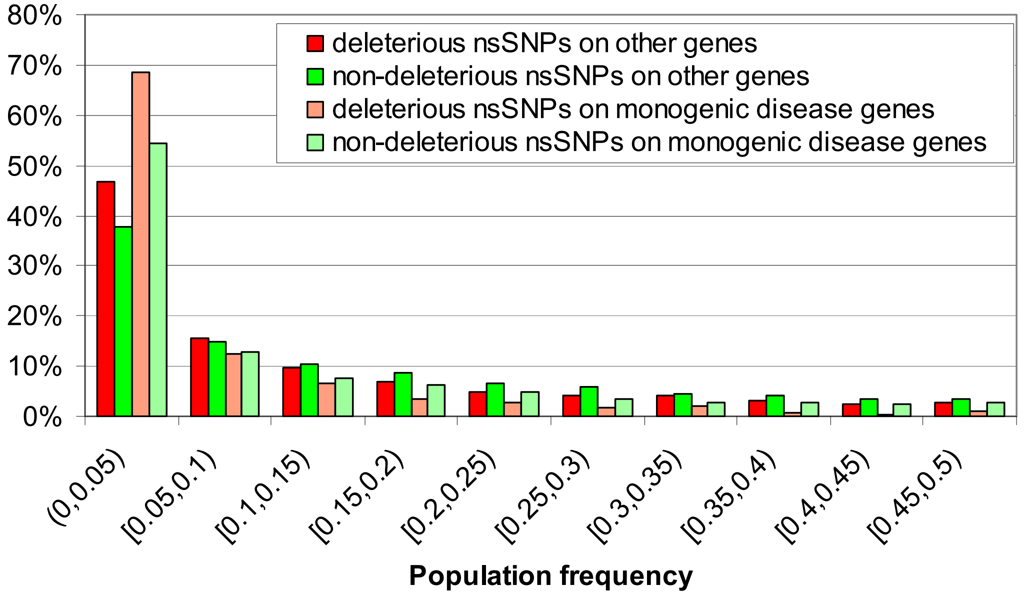


Figure 3-4. Distribution of SNP Frequencies in the Human Population.

Solid red bars show the fraction of all deleterious missense SNPs in each frequency range, for all non-monogenic disease genes. The hashed red bars show the same data for monogenic disease genes, Green bars show the corresponding data for Non-deleterious SNPs. As expected, low frequency SNPs are the most common, for all categories. Deleterious SNPs are biased towards low frequencies in both sets, but the effect is considerably stronger for monogenic disease genes.

Gene	SNP ID	SVM Stability	SVM Profile	Substit-ution	Freq	Source	Population	HGMD
CFTR	766874	-0.88	-1.75	S605F	0.002	Hapmap	afr,eur,chn,jap	
FCER1A	2298805	-0.73	-2.67	S101N	0.007	Perlegen	afr,eur,chn	
NTRK1	6336	-1.06	-1.17	H604Y	0.011	Hapmap	afr,eur,chn,jap	CM990977
DNASE1	1799891	-0.54	-0.77	P154A	0.011	Hapmap	afr,chn,jap	
CFTR	1800100	-1.06	-2.12	R668C	0.014	Perlegen	afr,eur,chn	CM950247
LYZ	1800973	-0.92	-0.72	T88N	0.015	Hapmap	afr,eur,chn,jap	
CHAT	8178990	-0.82	-1.26	L125F	0.021	Hapmap	afr,eur,chn,jap	
EPX	2302311	-1.32	-0.81	N572Y	0.027	Hapmap	afr,eur,chn,jap	
HFE	1800562	-1.00	-1.77	C194Y	0.028	Perlegen	afr,eur,chn	CM960828
TAP1	1057149	-0.80	-1.94	R708Q	0.029	Perlegen	afr,eur,chn	
CYP2A6	17791931	-0.81	-1.99	L160H	0.035	Perlegen	afr,eur,chn	CM980517
KLK3	17632542	-0.58	-1.67	I179T	0.036	Perlegen	afr,eur,chn	
PTGS2	5272	-1.40	-1.42	E488G	0.056	Hapmap	afr,eur,chn,jap	
HFE	1799945	-0.70	-0.66	H63D	0.085	Hapmap	afr,eur,chn,jap	CM960827
CYP2A6	5031017	-1.57	-1.61	G479V	0.125	Hapmap	afr	
OTOR	6135876	-0.91	-0.96	L31P	0.141	Perlegen	afr,eur,chn	
SERPINA7	1804495	-1.28	-1.38	L303F	0.203	Perlegen	afr,eur,chn	
AMACR	10941112	-1.51	-2.35	G175D	0.341	Hapmap	afr,eur,chn,jap	

Table 3-6. Very high confidence classifications of deleterious population SNPs in monogenic disease genes.

‘SVM Stability’ and ‘SVM Profile’ are the scores assigned by the two classification methods. A score < -0.5 is high confidence. The ‘Freq.’ column gives the mean frequency of each SNP over the populations. The ‘Population’ column lists the populations in which each SNP has been genotyped: afr: African, eur: European, chn: Chinese, jap: Japanese populations. Only five of these SNPs are in the HGMD database of disease causing mutations (IDs in the last column).

(Mori et al. 1990) (Waltz et al. 1990). These mutants alone are not sufficient to cause disease, since the resulting tendency for hyperthyroid is usually reversed by reduced thyroid hormone secretion. The high frequency is thus likely a consequence of a second factor being required for disease. There is no obvious explanation for the high frequency of the AMACR SNP.

Divergence Rates of Monogenic Disease-Associated Proteins

Figure 3-5 shows a comparison of divergence rates of monogenic disease proteins versus all others. A larger proportion of the most highly conserved proteins are non-disease, whereas at moderate to high conservation, a higher proportion is disease. At the lower conservation levels, non-disease proteins are slightly more common. This pattern can be rationalized as follows. Damage to the most conserved proteins is more likely to be lethal, and thus, not identified as disease causing. The lowest conserved proteins are likely buffered against deleterious changes in some way, and so are also not involved in monogenic disease. It is the more moderately to highly conserved genes where deleterious SNPs are likely to lead to disease, but not to be lethal. Other reports (Huang et al. 2004; Smith and Eyre-Walker 2003), using only average values, and separately analyzed Ks and Ka rates, come to contradictory conclusions. With more genomes becoming available, further study will be worthwhile.

Comparison with Mouse Knockout Data

The profile and stability models detect SNPs that reduce the level of protein function *in vivo*. The limit of reduced function is the absence of the gene. Thus, we

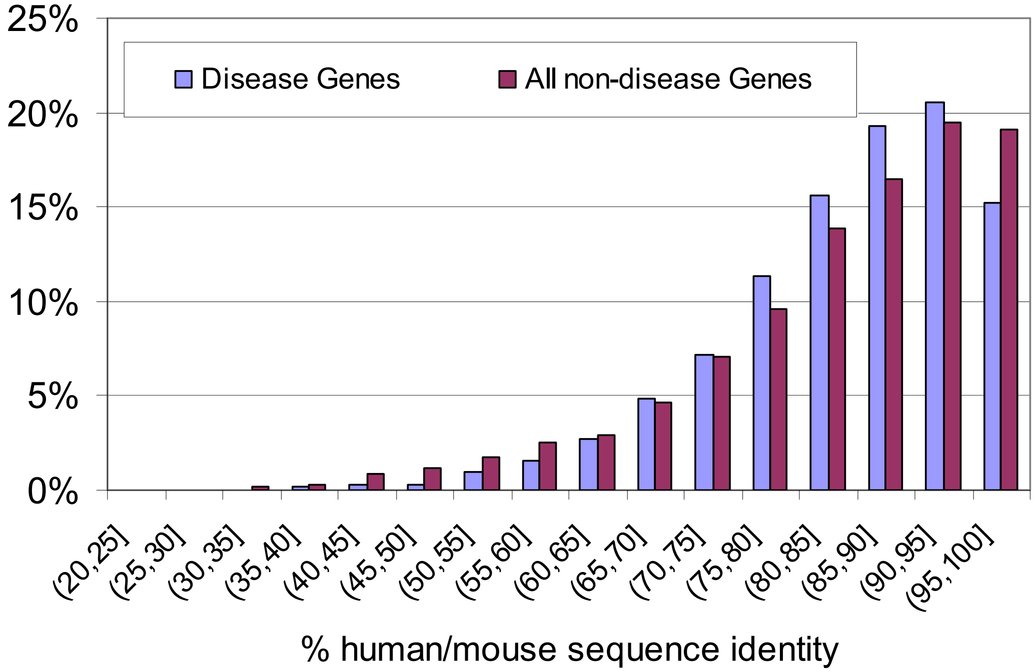


Figure 3-5: Protein Sequence Divergence Rates for Human Monogenic Disease Proteins (Blue bars) and all others (Purple bars).

Rates are expressed in terms of the sequence identity between each human protein and its mouse ortholog. Disease proteins have a larger proportion of high sequence identity mouse orthologs, showing that, on average, their sequences diverge more slowly than those of other proteins.

would expect a relationship between the response of the human phenotype to deleterious SNPs, and the response of mice to knockout of the corresponding orthologs. Mouse knockout data were obtained from http://www.bioscience.org/knockout/knochome.htm. In this database, genes are clustered into four knockout phenotype groups. The first group is of genes where the knockout is compatible with viability. This group is further subdivided into cases where there is a detectable effect on the phenotype, and cases where the phenotype is apparently unaffected. The other three groups are of genes where knockout causes post-, peri- and prenatal mortality.

Table 3-7 shows the fraction of monogenic disease genes found in each of the mouse knockout groups. The lowest fraction of monogenic genes is for the 'no effect' group of knockouts (8%), consistent with fully buffered genes generally not contributed to monogenic disease. The next lowest fraction is for the prenatal mortality set (28%), consistent with defects in these human genes probably resulting in a non-viable fetus, and so not be classified as disease associated. Approximately half of the other knockout groups have equivalent monogenic disease genes, consistent with non-lethal but significant impact on the phenotype often being classified as monogenic disease. Overall though, the correlations are not as high as might be expected. There are several possible reasons for that. As more mouse knockout data becomes available, a fuller analysis will be possible.

Frequency of Paralogs for Monogenic Disease and other Genes

A possible distinguishing feature between monogenic disease genes and the rest is that the phenotype is robust to reduced function on the latter because of

Phenotype		Total genes	Disease genes	Fraction DIsease
COMPATIBLE WITH VIABILITY	NO APPARENT EFFECT	13	1	8%
	With EFFECT	147	71	48%
POSTNATAL OR PERINATAL MORTALITY		51	22	43%
PRENATAL MORTALITY		29	8	28%

Table 3-7. Relationship between Mouse Knockout Phenotypes and Human Monogenic Disease Genes.

‘Total genes’ are the number of mouse knockouts in each phenotype category, and ‘Disease genes’ are the number for which the human ortholog is a monogenic disease gene.

redundancy of function – other genes can at least partly compensate for reduced activity. Full identification of possible substitute genes requires a detailed knowledge of human protein networks, not yet available. However, it might be expected that paralogs would often perform this role, and a number of such cases are known. For example, E-selectin and P-selectin are paralogous, with 40% protein sequence identity. Single gene knock-out mice show mild phenotypes, while the double knock-out mice have a severe disease phenotype, consistent with overlapping function (Frenette et al. 1996). On the other hand, there are many cases where paralogous genes are involved in different biological processes, for example malate and lactate dehydrogenases.

Paralogs were identified by searching each human protein sequence against all others, selecting relatives with a BLAST E-score of 10^{-3} or better. Table 3-8 shows the fraction of monogenic and other genes with at least one paralog. There is no difference between the two types of gene – in both cases about 87% have paralogs. We conclude from this that buffering mechanisms are more varied than just the use of paralogs.

	Monogenic Disease Genes		**Other Genes**	
	Count	%	Count	%
No paralogs	227	13%	705	13%
Paralogs	1460	87%	4887	87%

Table 3-8. Fraction of Monogenic and other genes that have paralogs.

Monogenic disease data are from HGMD (Stenson et al. 2003). ‘Other Genes’ are other human genes containing at least one SNP classified as deleterious. There is no difference in the fraction with paralogs for the two sets, suggesting other mechanisms dominant in shielding the phenotype from the adverse effects of deleterious SNPs in the non-monogenic disease genes.

Discussion & Conclusion

The main conclusion of this study is that about one quarter of the known missense SNPs in the human population are significantly deleterious to protein function *in vivo*. Others have reported a figure of about 1/3 (Chasman and Adams 2001; Ramensky et al. 2002). It has also been suggested that the fraction is much lower (Ng and Henikoff 2003), with false positives, errors in dbSNP, and known monogenic disease mutations inflating the apparent value. We have carefully controlled for false positives and false negatives in two ways. First, using two independent SNP classification methods has allowed us to check that the expected error levels are obtained for the combined methods, so validating the individual values. Second, we have calculated the apparent deleterious rate taking into account the error levels, and obtained a best fit for the underlying true deleterious rate. We have also examined the difference in apparent deleterious rate for all of dbSNP and a validated subset. There is indeed a higher value of about 1/3 for all dbSNP, but the value of a quarter is obtained on reliable data. Some of the deleterious SNPs are in known monogenic disease genes, but about 80% of the dbSNP ones, and 70% of the validated set, are not.

Some of the new deleterious SNPs in monogenic disease genes are candidates for previously unrecognized disease causes. The deleterious SNPs in non-monogenic disease genes are candidates for contributing to complex disease traits. Presumably, the network environment of the proteins concerned buffers the effect on the phenotype. This view is supported by the analysis of the relationship between monogenic disease genes and mouse knockout phenotypes – knockouts with intermediate impact on the phenotype are more likely to be orthologs of human monogenic disease genes. A simple form of buffering is overlapping function with paralogous proteins. For example, a T cell mediated immune response will involve many different T-cell receptors. We have found deleterious SNPs in some of these proteins (Wang and Moult 2003), but redundancy through paralogs will provide buffering. Surprisingly, we did not find that monogenic disease genes are less likely to have paralogs than others, so this mechanism is probably only one of a number. A proper understanding these buffering processes will require a detailed knowledge of the relationship protein function and network behavior.

Many of the deleterious SNPs in non-monogenic disease genes are relatively rare. In one sense, this is expected, since overall, there are many more rare SNPs than common ones. The low frequency of deleterious SNPs may contribute to relatively rare complex traits, or they may contribute in many combinations to produce common traits (Smith and Lusis 2002) (Pritchard and Cox 2002).

For complex diseases, variation in a single gene only marginally increases risk, and as a consequence, most association studies present weak and sometimes inconsistent results (Prince et al. 2001). The deleterious SNPs found in this and other analyses provide additional information that can be used to select SNPs for inclusion in association studies, or, in larger scale studies, to provide prior probabilities that can be incorporated into the statistical model.

The analysis of human SNPs was done using a structure based method (Yue et al. 2005), and a sequence profile based method. The sequence method has a larger coverage of missense SNPs because it does not require knowledge of three dimensional structure. Also, since sequence methods are based on evolutionary selection information extracted from multiple sequence alignments, they are not limited by current knowledge of protein function and structure, and so include a wider range of effects. On the other hand, the sequence method assumes that deleterious SNPs will eventually be removed during evolution. While this assumption may be true for those genes associated with monogenic disease or serving as major contributors to complex diseases, it may not be as true for those with only subtle effects on the phenotype. For this reason, it is desirable to develop broadly based mechanistic models of SNP impact.

Methods

Construction of the Deleterious Variant Dataset and Non-deleterious Variant Dataset

The same two datasets were also used for the stability model as described in Chapter 2. Please refer to page 44 for the procedures of data construction.

Source of Human Population Missense SNPS

SNPs were obtained from NCBI dbSNP, build 124. Many of the dbSNP entries are not verified (are based on single observations, or population frequency data have not been deposited). A confirmed SNP set was built from data in Perlegen (as of May 2005) and the Haplotype genotyping projects (Phase I, as of May 2005). Files containing SNP and frequency information were downloaded from Perlegen and Hapmap project websites (http://genome.perlegen.com and http://www.hapmap.org/). These two datasets was processed as follows: 1) Both datasets were mapped to dbSNP RefSNP clusters. Hapmap provides a link from each record to a RefSNP ID; the Perlegen submission SNP ID and the mapping table, SNPSubSNPLink, between submission SNP IDs and RefSNP IDs in dbSNP build 124 were used to link each Perlegen record to the related RefSNP cluster; 2) For each RefSNP entry, mean frequencies were calculated from the three Perlegen populations, and from the available Hapmap populations; 3) In cases where data are available for both sources, the Hapmap information was discarded. dbSNP links were used to map each SNP to the corresponding amino acid substitution.

Construction of Sequence Profiles

Each human protein sequence was searched against the NR (Non-redundant Protein Database) using PSIBLAST (Altschul et al. 1997) with an E-score cutoff of 10^{-3} and three search rounds. The PSIBLAST sequence alignment (profile) and the position specific scoring matrix (PSSM) were retained for further use. Profiles were filtered as follows:

1. Closely-related proteins were removed: If a pair of proteins has more than 90% sequence identity in PSIBLAST, one was eliminated from the profile.

2. Less reliably aligned proteins were removed: Any protein with less than 30% sequence identity to the query human sequence was removed.

3. Regions of the alignment where more than 50% of the sequences have a gap were removed.

Features for the Support Vector Machine

The following five features were used for the SVM:

1. The probability of substituting the variant residue type 'a' at position 'j' in the sequence alignment, P(a,j), taken from the corresponding matrix element in the PSSM.

2. The Entropy at each position 'j' in the alignment is calculated using the Shannon entropy formula (Shannon, C.E. A Mathematical Theory of Communication. The Bell Systems Technical Journal, 27 (1948), 379-423):

$$S_j = -\sum P_i \log_2 P_i$$

Where the sum is over the twenty possible amino acids, and P_i is the probability of particular residue type 'i' at this position. Probabilities are calculated from the filtered alignment profile.

3. The mean entropy <S> over the sequence is calculated by averaging over all sequence positions.

4. The standard deviation of the entropy over all positions is calculated as:

$$\sigma(S) = [(\sum_i(S_i - \langle S \rangle)^2)/(N-1)]^{1/2}$$

Where the sum is over all sequence positions, and S_i is the entropy at a particular position, and N is number of sequence positions.

5. The entropy at each position j is expressed as a Z score:

$$Z_j = (S_j - \langle S \rangle)/ \sigma(S)$$

Support Vector Machine (SVM)

The five parameters described above: probability of accepting that amino acid substitution, entropy, mean entropy, standard deviation of the entropy and the entropy Z score, were used as features to train a SVM. The deleterious variant set consisted of these values for all the monogenic disease causing residue positions, and the control set were the values for the inter-species amino acid differences. $\text{SVM}^{\text{light}}$ (http://svmlight.joachims.org/), an implementation of SVM in C, was used, with a linear kernel. Weights were assigned to the disease and control data sets to compensate for their different sizes, such that they contributed equally to determining the partitioning surface. 40% of each of the two sets was randomly selected to train the SVM. The remaining 60% were used to evaluate accuracy. The training and testing procedure was repeated 30 times. For each trial, the false negative rate (the fraction of deleterious variations mis-classified as non-deleterious) and false positive rate (the fraction of non-deleterious variations mis-classified as deleterious) in the test dataset were calculated. The average false positive and false negative rates provide the measure of the classification accuracy. The distance of a data point from the partitioning surface provides an approximate measure of confidence in a classification.

Calculation of the Expected Sensitivity and Specificity of the Combined Stability and Profile Methods

Assuming the two methods are independent: For sensitivity, if $P_1(T)$ represents the probability of identifying a true positive for model 1 and $P_2(T)$ represents the corresponding value for model 2, the probability that both models identify the same true positive is $P_{12}(T) = P_1(T).P_2(T)$. For specificity, if the probability of model 1 producing a false positive is $P_1(F)$ and for model 2 is $P_2(F)$, the probability that both models identify the same false positive is $P_{12}(F) = P_1(F).P_2(F)$. The expected specificity is then $1 - P_{12}(F)$.

Estimate of Protein Divergence Rate from Human and Mouse Orthologous Genes

Mouse orthologs were taken from the NCBI HomoloGene database (Wheeler et al. 2005). For each orthologous pair, the BLAST sequence identity was calculated between the all refseq mouse protein sequences and those of all the corresponding human refseq entries, and the highest value was used. (This procedure is necessary, since each gene may have multiple protein isoforms).

Matching of Mouse Knockouts with Human Genes

The OMIM ID of each available mouse knockout gene was extracted and matched to the NCBI locuslink database, to identify the corresponding human gene name. Human curation was used to match remaining mouse genes and verify each link. The matched human genes were compared to those in the HGMD database, to find the subset involved in monogenic disease.

Chapter 4: SNPs3D: Candidate Gene and SNP selection for Association Studies

Introduction

Much of our present knowledge of the relationship between genotype and disease comes from statistical studies of the correlation between particular genetic variants and the likelihood of a specific disease. Linkage analysis, which tracks the transmission pattern of genetic markers within a pedigree family, has been successful in identifying over one thousand human monogenic disease genes (Stenson et al. 2003). On the other hand, there has so far been less success with common human diseases, such as hypertension, Alzheimer's, asthma and cancer. Susceptibility to these is affected by multiple genes, as well as environmental factors. The risk from any single genetic variant is low, so that linkage analysis sample sizes are usually too small to provide statistically significant disease/genotype relationships. Association studies, based on analysis of genetic differences, particularly SNPs, between those with and without a disease in a broader population, are more powerful for detecting such low signals. Approximately 10 million human SNPs have so far been identified (Sherry et al. 2001). Currently, association studies depend on choosing a subset of these which includes those influencing the probability of disease, or that are in linkage disequilibrium with those that do so. A primary purpose of the SNPs3D resource is to provide a means of selecting candidate genes likely to influence disease susceptibility, and to further select the most relevant non-synonymous SNPs within those genes.

Rapid accumulation of new data on human SNPs, knowledge of the complete human genome sequence, and increasing information on biomarcomolecular interactions is opening the way to a more mechanism based understanding of the relationship between genotype and disease. At present, the relevant information is still very incomplete, and is scattered across many databases and thousands of articles. A second primary purpose of the resource is to collect and integrate as much as possible of the molecular level data relevant to the mechanisms that link genetic variation and disease.

To achieve these goals, the resource is organized into three modules. One module generates lists of candidate genes for any specified disease, based on an analysis of the relationship between the disease and genes, as reflected in the literature. The second module provides a interactive graphical gene-gene network, built from literature associations, known protein-protein interactions (Bader et al. 2003)(http://bind.ca/), and existing pathways (Kanehisa et al. 2004) (http://www.genome.jp/kegg/). The third module provides information on the relationship between non-synonymous SNPs and protein function.

The identification of candidate genes and construction of gene networks both make use of simple text mining techniques. Concept profiles are constructed for each disease and for each gene. Each concept (a disease or a gene) is represented by an ordered list of words and terms most closely associated with the concept. The set of words and terms is complied from the contents of the approximately 80,000 PubMed abstracts (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=PubMed) that have been manually associated with one or more human genes in the NCBI Gene database (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene), using natural language processing (http://www.lsi.upc.es/~nlp/SVMTool). Pairs of concepts, such as two genes or a disease and a gene, are linked by the overlap of their keyterm profiles. We call the resulting gene-gene network a KnowledgeNet, since it is derived directly from knowledge in the literature. Only two types of concept, gene and disease, are discussed in this chapter. However, the KnowledgeNet can also be used in others ways, for example investigating the relationship between a biological process (e.g. glycolysis) and genes.

In SNPs3D, the likely functional impact of non-synonymous SNPs is assessed using two previously developed methods (Wang and Moult 2001; Yue et al. 2005; Yue 2005). One method makes use of protein structure to identify which amino acid substitutions significantly destabilize the folded state. The results show that up to three-quarters of monogenic disease single residue mutants act in that way (Yue et al. 2005). The second method identifies deleterious substitutions through analysis of the extent and nature of amino acid conservation at the affected sequence position (Yue 2005). Access to details of both analyzes is provided through the web interface. Links to another publicly available non-synonymous SNP analysis tool are also provided (Dantzer et al. 2005) (http://mutdb.org/).

SNPs3D aims at integrating all of the available data relevant for assessing the likely role of particular genes and SNPs in a disease. The emphasis is on providing the users access to as much of the underlying information as possible, so that they may make informed judgments. To this end, in addition to SNP impact analysis, links are provided to relevant abstracts, the GAD (The Genetic Association Database)

(Becker et al. 2004) (http://geneticassociationdb.nih.gov/), OMIM (Hamosh et al. 2005) (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM) and HGMD (Stenson et al. 2003) (http://www.hgmd.org/) disease databases, GO annotation (Harris et al. 2004) (http://www.geneontology.org/), expression profile data (Su et al. 2002), and mouse knockout results (http://www.bioscience.org/knockout/knochome.htm). Data are updated regularly. Exploration of gene networks and access is to all information is facilitated by a Java based graphical interface.

Results

Analysis of SNPs in each Human Gene

A primary function of the SNPs3D resource is to provide a way of identifying those non-synonymous SNPs that are likely to have a deleterious impact on molecular function *in vivo*, so these may be included in association studies. An analysis of the likely functional impact of all human non-synonymous single base variants in the HGMD (as of 02/09/2002 , 9,625 variants in 696 genes) (Stenson et al. 2003) and dbSNP (Build 124, 29,485 SNPs in 11,303 genes) databases (Sherry et al. 2001) (http://www.ncbi.nlm.nih.gov/projects/SNP/) is provided, using the previously developed methods (Yue et al. 2005; Yue 2005). Links to another available analysis (Dantzer et al. 2005) (http://mutdb.org/) are also included. The analysis is organized by gene. The structure/stability method (Wang and Moult 2001; (Yue et al. 2005)) requires knowledge of structure. Availability of experimental structures or sufficiently accurate structure models limits coverage to about 37% of monogenic disease variants in HGMD and 10% of variants in dbSNP. Greater availability of sequence information compared to structure allows a much higher fraction of variants to be analyzed (92% and 57% HGMD and dpSNP respectively) with the sequence profile method.

Both methods make use of a machine learning technique, the support vector machine (SVM), to assign each SNP as deleterious or non-deleterious to protein function. The SVM is trained on monogenic disease data, so that the definition of deleterious is ‘sufficiently damaging to protein function *in vivo* as to be consistent with a monogenic disease outcome'. Benchmarking has yielded false positive and false negative rates of 15% and 26% for the stability method and 10% and 20% for the sequence profile method. The higher false negative rate for the stability method reflects the fact that only stability effects on *in vivo* function are included. Approximately 30% of the non-synonymous SNPs in dbSNP are assigned as deleterious. Very few of the dbSNP cases are known to be associated with monogenic disease, and so most the deleterious ones are candidates for contributing to complex disease traits. As illustrated later, in many cases, low impact on the phenotype is likely the result of network level buffering against loss of function for individual proteins.

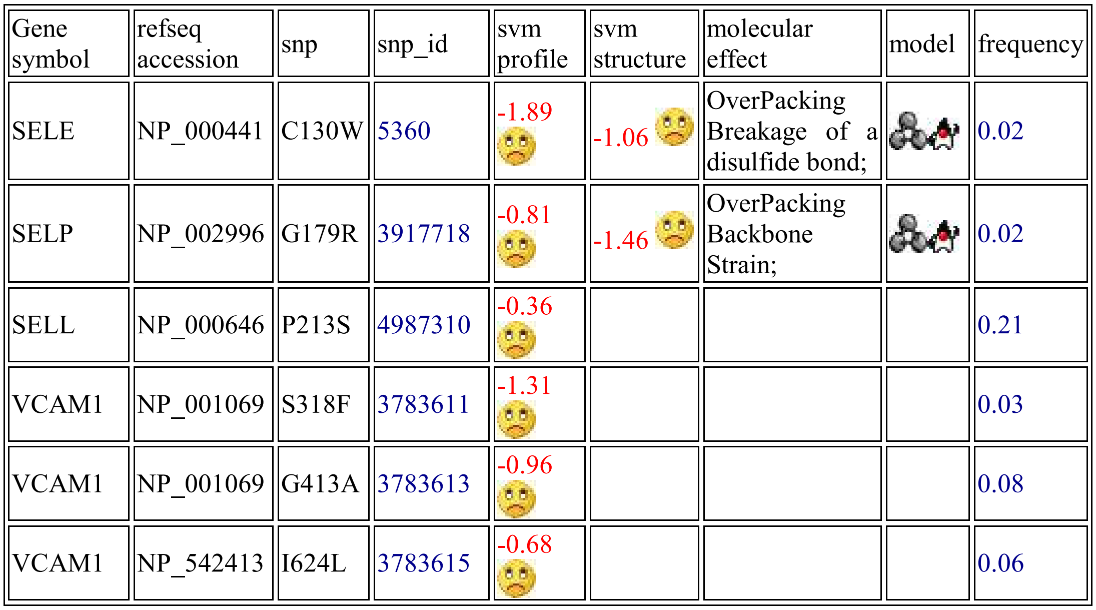
Details of the analysis of each SNP are provided on additional pages. For the profile model, a user can inspect the multiple protein sequence alignment from which the result is derived. For the structure/stability model, feature values (for example, surface accessibility, electrostatic interactions and hydrophobicity) are provided, as well as an interactive molecular graphics interface (powered by Jmol, http://jmol.sourceforge.net/) displaying the affected residue in its three dimensional structural context.

An example of Deleterious SNP Analysis

To illustrate the SNP analysis process, we consider SNPs in the selectins, proteins involved in the early inflammatory response, playing a role in the accumulation of blood leukocytes at sites of inflammation. SNP analysis for relevant genes may be accessed by typing a disease or process name into the corresponding search window. Entering 'inflammation' returns a ranked list of genes with abstracts containing that term, hyperlinked to their SNP analysis pages. Entering a more specific search term, such as ‘selectin’ returns a list of relevant genes, including the members of the selectin family SELE, SELP and SELL, as well as proteins they interact with. Entering a specific gene name, such as SELE, takes the user directly to the analysis of SNPs in that protein. Table 4-1 shows a composite of the screen information for some inflammation related SNPs in selectins E, P and L and VCAM1. Each of these SNPs is classified as deleterious by the sequence profile method (indicated by the negative SVM scores). The SNPs in SELE (C130W) and SELP (G179R) are also analyzed by the structure/stability model, and are found to be deleterious by this criterion as well (a disulfide bridge is broken in SELE, there is overpacking and backbone strain in SELP). As discussed below, further insight into the relationship between these SNPs and the inflammatory response is provided by consideration of the inter-gene relationships.

Gene-Gene Relationships

Concept profile overlaps were used to score the relationship between all pairs of human genes in the current NCBI Entrez Gene database. Table 4-2 shows part of the resulting gene-gene relationship matrix, involving hypertension genes. Angiotensin-converting enzyme (ACE) and angiotensinogen (AGT) share 96 specific keyterms, such as ‘sodium intake’, ‘renin-angiotensin-system’ and ‘blood pressure’; generating a very strong (43.8) link between them. Many of the shared keywords also have relatively high weights. (That is, the frequency is high in abstracts for these genes, compared with all abstracts). In contrast, the link between ACE and arginine vasopressin (AVP) is much weaker, with a score of 0.8, (still above the average for



Gene symbol	refseq accession	snp	snp_id	svm profile	svm structure	molecular effect	model	frequency
SELE	NP_000441	C130W	5360	-1.89 ☹	-1.06 ☹	OverPacking Breakage of a disulfide bond;		0.02
SELP	NP_002996	G179R	3917718	-0.81 ☹	-1.46 ☹	OverPacking Backbone Strain;		0.02
SELL	NP_000646	P213S	4987310	-0.36 ☹				0.21
VCAM1	NP_001069	S318F	3783611	-1.31 ☹				0.03
VCAM1	NP_001069	G413A	3783613	-0.96 ☹				0.08
VCAM1	NP_542413	I624L	3783615	-0.68 ☹				0.06

Table 4-1. Example interface page of candidate SNPs for inflammation related disease.

Two support vector machine (SVM) models, based on sequence profiles (Yue 2005) and structural stability (Yue et al. 2005) are used to analyze SNPs in candidate genes for inflammation. SNPs are classified as deleterious (negative SVM score) or not to protein function *in vivo*. SNP population frequency information is extracted from the NCBI dbSNP database.

	ACE	AGT	AVP	…
ACE		43.8	0.8	…
AGT	43.8		0.4	…
AVP	0.8	0.4		…
..	…	…	…	

Table 4-2. Subsection of the KnowledgeNet gene-gene linkage matrix.

All three genes are associated with blood pressure regulation. ACE and AGT are strongly linked, other links are near the average value of 0.5.

non-zero relationships in the matrix, which is 0.5). There are only two shared keyterms between these genes: 'polydipsia’ and ‘hypotension’. ‘Hypotension’ represents a true concept overlap between these two genes, since both are involved in the regulation of blood pressure. ‘Polydipsia’ is a symptom found in more than one disease. One of these is Autosomal dominant familial neurohypophyseal diabetes insipidus (ADNDI), some times caused by a missense mutation in AVP (Smith et al. 2002). Mutations in ACE have also been shown to be a risk factor in a different disease, schizophrenia, for which polydipsia is also a symptom (Shinkai et al. 2003). Thus linkage of ACE and AVP through this term is a not a consequence of their joint role in blood pressure regulation. These indirect linkages are a source of noise in the matrix, but are generally rare.

Figure 4-1 shows that the distribution of scores between gene pairs has an approximately power law distribution, with many scores near the minimum of 0.001, and a few high scores of up to 300. Pairs of genes which are in the same KEGG pathway (Kanehisa et al. 2004) tend to have a stronger link than others, with median and mean scores of 0.5 and 2.5, while for all genes the corresponding values of 0.2 and 0.5 respectively. When only those pairs of genes involved in physical interactions included in the BIND database (Alfarano et al. 2005) are considered, the median and mean are dramatically higher, at 3.2 and 9.0 respectively. Note that it is not our aim to reproduce either of these known gene-gene relationships, but to introduce a more general, literature based measure.

Figure 4-2 shows the distributions of the number of gene links, for monogenic disease (defined by inclusion in the HGMD database (Stenson et al. 2003) and all

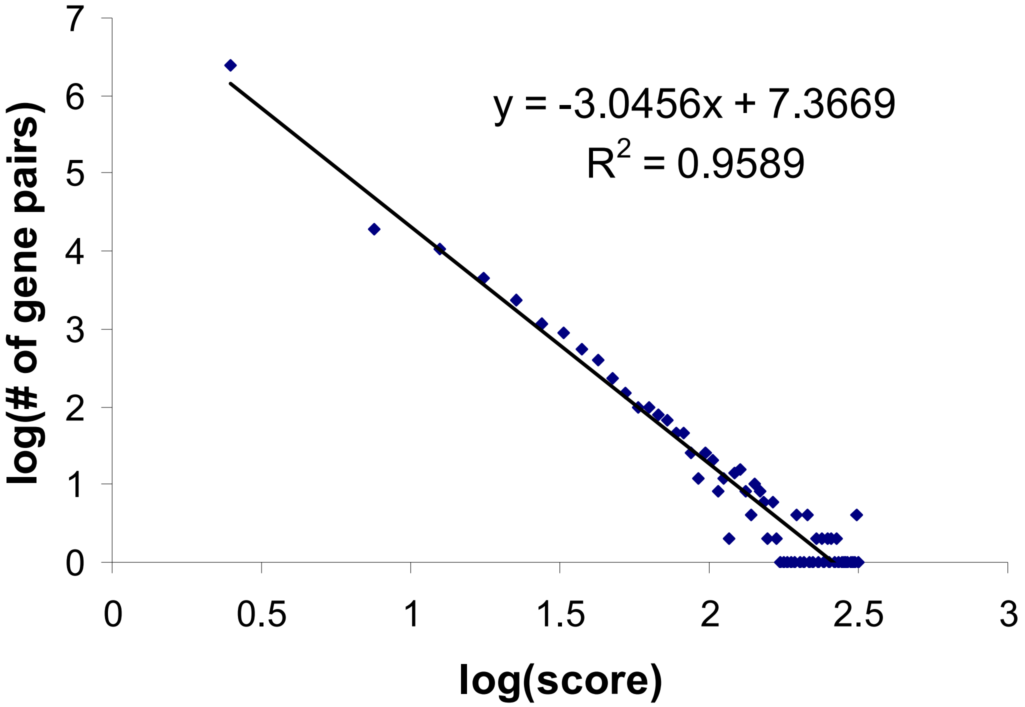


Figure 4-1. Log- log plot of linkage scores in the gene-gene KnowledgeNet. Scores follow an approximately power law distribution, with a few very high scoring relationships (up to a value of 300), and many relatively weak ones.

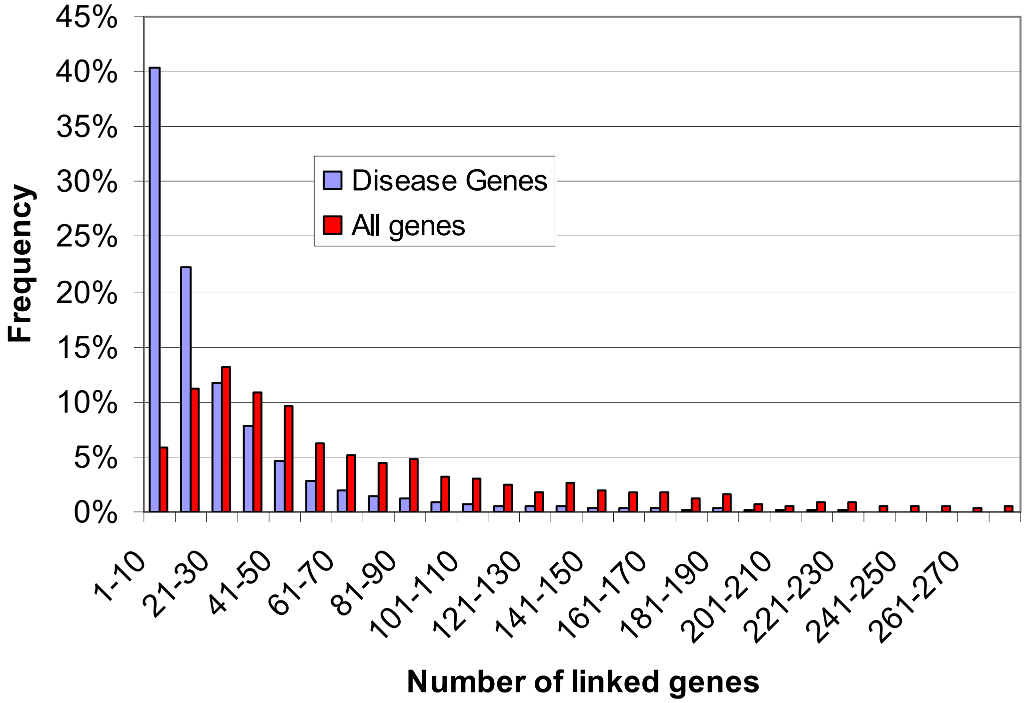


Figure 4-2. Distribution of the number of links to each gene in the gene-gene KnowledgeNet. Blue bars show the distribution for all genes with at least one link (15,799) and red, the distribution for 1669 linked HGMD monogenic disease genes. The tail is truncated – the highest linkage is 493, for TP53. Genes with no interactions above the threshold score of 0.5 are not included.

genes. Disease genes tend to be linked to more genes than non-disease genes, reflecting the fact that they are usually well studied, and have been placed in a network context.

Using the Gene-Gene KnowledgeNet to Investigate SNP-Phenotype Relationships

The SNPs in table 4-1 are classified as significantly deleterious to protein function, and are in genes involved in the inflammatory response. However, none of these SNPs is known to produce a disease phenotype. We next illustrate how the KnowledgeNet can be used to investigate the complex relationships between the effect of these SNPs on protein function and the disease phenotype, through network level buffering against defective protein components. For simplicity, we consider one pair of genes with deleterious SNPs, selectin E and selectin P. The sidebar on the SNP analysis page provides direct access to a wide range of information relevant to this question, including OMIM, pathways, GO annotation, mouse knockout results, and tissue specific expression data, and relevant abstracts. Clicking ‘Gene Graph’ in the left sidebar creates a Java window displaying the gene-gene relationships centered on SELE.

A large amount of information is accessible through the Java interface. At the moment, we are specifically interested in possible buffering mechanisms that shield the phenotype from these deleterious SNPs. One such buffering mechanism is overlapping protein function, and many proteins with overlapping function are homologous (Kafri et al. 2005) . Right clicking on the E-selectin node triggers a popup menu, including an option for highlighting all sequence homologs of that node in the graph. L-selectin and P-selectin are seen to be homologous to E-selectin, suggesting possible functional redundancy. The redundancy of selectins E and P is supported by the information obtained from the mouse knockout link in the same menu, which reveals that single mouse knockouts of each gene produce a mild phenotype, while the double knockout is severe (Frenette et al. 1996). Further support is provided by inspection of the expression profiles for the selectins, which shows a similar tissue specific pattern for Selectin E and selection P, with significant expression in multiple tissues, while selectin L is found in only a few tissues. Thus, an individual homozygous in either one of the deleterious SNPs will likely have a subclinically affected inflammatory response, because of redundancy of function. But an individual with both may have an epistatic interaction between them, and be seriously sick. Both are candidates for inflammation related disease association studies.

Candidate Gene Lists for Diseases

As discussed in the Introduction, the candidate gene approach is still widely used in association studies. Since knowledge of complex diseases is limited, a comprehensive list of candidate genes and a method of ranking those genes by their disease-relevance is important in designing a good association study. The 'Disease Candidate Genes' module is used to list and rank candidate genes by building a concept profile for the disease and comparing it with the profiles for each human gene. The resulting ranked list of candidate genes can be edited by the user, before further analysis. The Java graphical interface provides access to the resulting gene network, helping a user navigate through the relationships and associated data.

We have pre-complied candidate genes lists for a set 76 diseases, taken from the NCBI on-line book, 'Genes and Disease' (http://www.ncbi.nlm.nih.gov/books /bv.fcgi?rid=gnd). A list for any additional disease may be generated by entering the disease name in the web interface.

Table 4-3 lists the 16 diseases associated with the most genes, using an association threshold of 0.05. (Disease-gene profile overlaps have scores ranging from 0 to 24.5 with a mean of 0.04). Figure 4-3 shows the distribution of the number of genes using this threshold. Cancers tend to have the largest number of candidate genes, with the highest value of 197 genes for lung cancer. Next ranking are well studied common diseases such as asthma, hypertension, inflammation, obesity, Alzheimer's disease, epilepsy, atherosclerosis and deafness. The number of genes associated with a particular disease primarily reflects the complexity of phenotype, but may also partly reflect the current state of knowledge. Not surprisingly, nominally monogenic diseases tend to have the least number of candidate genes. However, these are often not monogenic in this analysis. For example, Phenylketonuria (PKU) has 14 associated genes. As expected, in this case the primary disease gene (PAH - phenylalanine hydroxylase) has a very high linkage to the disease, with a score of 23, while all other genes have scores less than 0.5. The web resource provides a ranked list of candidate genes for each disease.

In all, 2,582 genes are associated with one or more of the 76 pre-complied diseases, using a threshold score of 0.05. TP53 is associated with the most diseases (23). The number of diseases a gene is associated with increases with the number articles associated with that gene.

Disease	Score >0.05
Lung Cancer	197
Prostate cancer	190
Gastric Cancer	142
Pancreatic Cancer	134
Breast Cancer	133
Diabetes Mellitus	130
Asthma	124
Retinoblastoma	116
hypertension	113
Bladder Cancer	109
Epilepsy	107
Inflammation Related	107
Atherosclerosis	99
Alzheimer Disease	99
Deafness	94
Cervical Cancer	93

Table 4-3. Diseases with the largest number of significantly associated candidate genes.
Cancers tend to have the largest number of candidates, followed by common complex trait diseases.

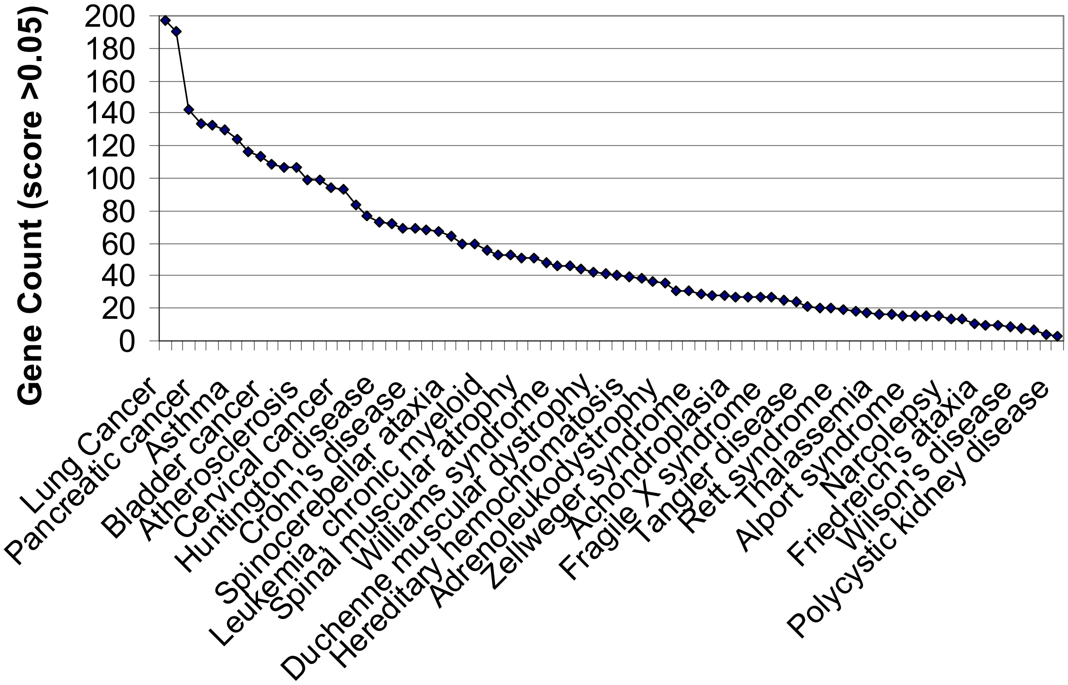


Figure 4-3. Distribution of the number of candidate genes for a set of 76 diseases. The curve shows the distribution using a disease-gene linkage threshold of 0.05. Cancers and common human diseases tend to have many candidate genes, but monogenic diseases typically have more than one candidate as well.

KnowledgeNet Analysis of Candidate Genes and SNPs

Once a candidate gene list is available, it is useful to be able to efficiently access the underlying literature, and to generate a list of deleterious SNPs in the genes of most interest. As an example of this process, we consider one of the pre-built disease candidate lists, for hypertension. Clicking on the disease returns a list of the candidate genes, ranked by confidence of disease relevance, based on profile overlap with the disease. Table 4-4 shows the top part of the list. Highest ranked are well known hypertension-related genes, for example, angiotensinogen (AGT) and angiotensin I converting enzyme (ACE). Each gene in the list is linked directly to local copies of the relevant abstracts, with color highlighting of appropriate words, so that a user may very rapidly assess the evidence for candidate status. There are also links to OMIM (Hamosh et al. 2005) and the NIA genetic association database information (Becker et al. 2004), providing sources of expert information on disease relevance.

Since hypertension is a complex trait, with susceptibility related to SNPs in multiple genes as well as the interactions between them, the ability to navigate the network of candidate genes is an important facility of the resource. Viewing the set of candidate genes in the Java graphical interface provides the mechanism for this. Figure 4-4 shows a screen snapshot of the graphical interface for the hypertension candidate gene network. Strongly associated genes cluster in the display. In particular, in this case, the four primary blood pressure regulation pathways form distinct groups, indicated by the black ovals. Among these, the renin-angiotensin pathway (A), controlling absorption of sodium, is the most studied, and most of its

Gene Symbol	Candidate SNPs	OMIM	GAD
AGT	1	Y	N3/Y19
ACE	6		N6/Y24
AGTR1	2	Y	Y11
GNB3	2	Y	N1/Y6
HSD11B2	1		Y1
CYP11B2	2		N1/Y2
BMPR2	0		
ADD1	1	Y	N5/Y4
REN	3		Y3
EDN1	0		

Table 4-4. Top ranking candidate genes for hypertension.

The list was complied on the basis of the overlap of the disease concept profile with those of the individual genes. ‘Candidate SNPs’ shows the number SNPs classified as deleterious in each gene. The ’OMIM’ column indicates which genes are associated with essential hypertension in that database. The ‘GAD’ column shows the number of votes for or against a role for each gene in hypertension in the Genetic Association Database (Becker et al. 2004).

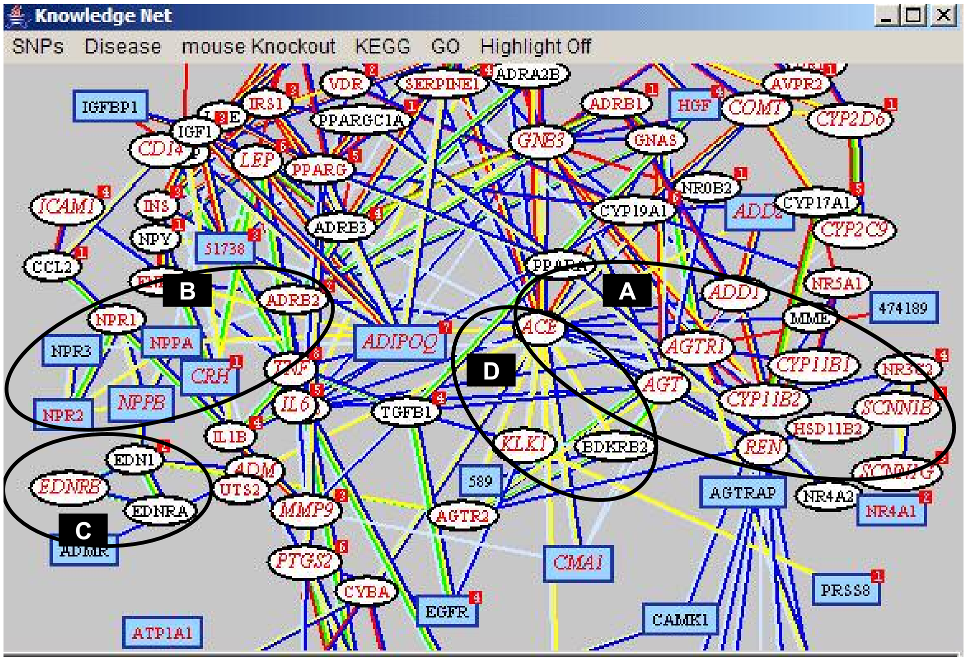


Figure 4-4. Graphical Interface for the KnowledgeNet of candidate genes for hypertension.

The four larger ovals circle the clusters of genes in each of the primary blood pressure regulation pathways. Oval symbols are used for genes involved in monogenic disease, rectangular symbols for the rest. Red indicates that one or more population SNPs are classified as harmful at the molecular level. Italic red text indicates that one or more population SNPs with population frequency information are predicted to be deleterious.

The length and color of the edges represent the strength of the link between pairs of genes. Red edges link genes sharing the same abstracts. Short edges link genes sharing a large number of biological keywords.

Subsets of nodes can be highlighted by a number of criteria, such as membership of the same KEGG pathway, or homology, or SNP frequency.

genes have been implicated in monogenic types of hypertension (indicated by the oval gene symbols). The other pathways all influence blood pressure through vascular constriction via: (B), regulation by endothelin (EDN1); (C), regulation of natruretic peptide (NPPA, NPPB, NPPC); and (D), the bradykinin-killikrien pathway. Figure 4-5 shows a simplified version of the pathways and their inter-relationships, derived from browsing the interface, reviews (Lifton et al. 2001) (Turner and Boerwinkle 2003), and on-line data (http://www.cvphysiology.com/Blood%20Pressure /BP001.htm). The pathways are highly interconnected. For example, both natruretic peptide and bradykinin also act as antagonists of the rennin-angiotensin pathway, and are able to relax vascular contraction and down-regulate blood pressure. Conversely, ACE, which activates AGT in the renin-angiotensin pathway, can inactivate bradykinin.

This gene/disease network for hypertension provides a number of deleterious SNPs for association studies. A sample of these is shown in table 4-5. All are classified as deleterious to protein function by the sequence profile method and the structure/stability method. The first is R333W in rennin, which results in the loss of salt bridge and thus is likely to cause loss of function. Given rennin’s role an up-regulator of blood pressure, this SNP is a candidate for involvement in hypotension. The second SNP, I444T, occurs in the hydrophobic core of angiotensin-converting enzyme (ACE) and causes a large loss of buried hydrophobic area. ACE is in the same pathway as rennin, and has an established role in blood pressure related disease. Mutants of ACE have been associated with monogenic-type hypertension (O'Donnell et al. 1998), and ACE knockout mice show ‘subnormal blood pressure, kidney

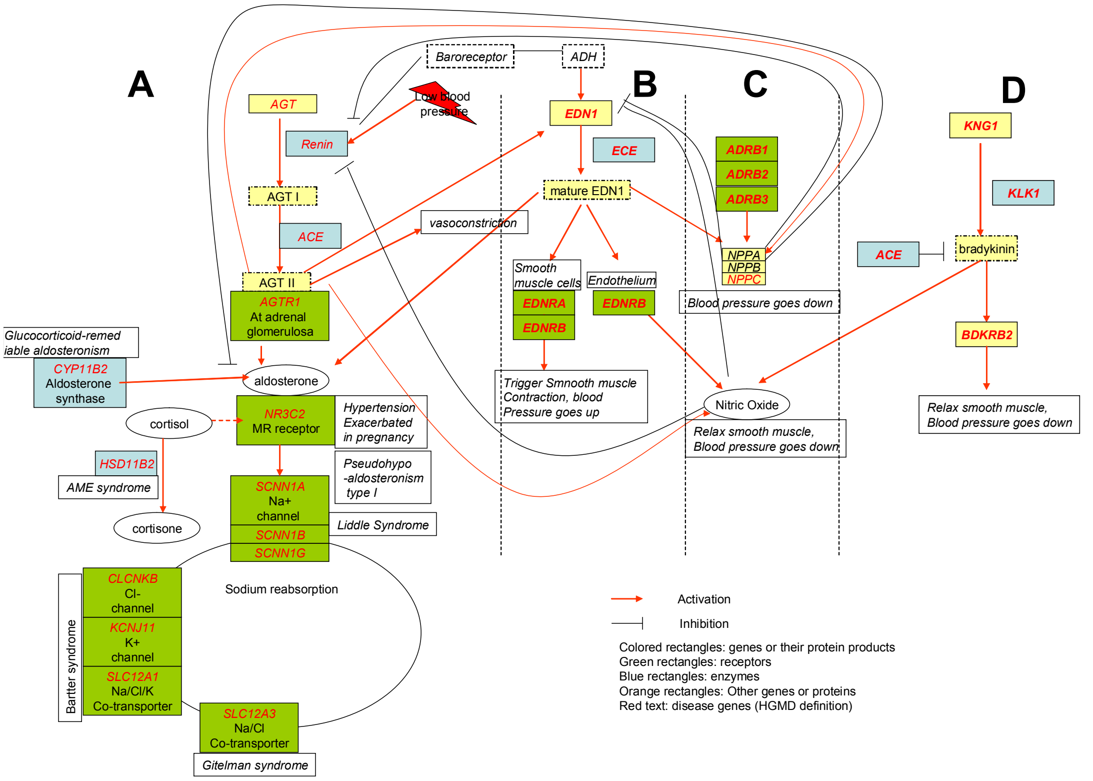


Figure 4-5. Simplified view of the four primary candidate pathways involved in hypertension.

A: renin-angiotensin pathway; B: regulation by endothelin (EDN1); C: regulation by natruretic peptide (NPPA, NPPB, NPPC); D: the bradykinin-killikrien pathway.

RefSNP ID	Gene Name	Refseq Protein	SNP	SVM profile	SVM structure	Structure and Sequence Properties	dbSNP ID and Population Frequency
rs11571098	REN	NP_000528	R33W	-1.63	-0.21	Salt Bridge lost	ss20420843:4% (African American)
rs4976	ACE	NP_690044	I444T	-1.26	-1.15	Hydrophobic Interaction loss	ss6413:5% (Multination) (Halushka et al. 1999)
rs5247	CMA1	NP_001827	H66R	-2.51	-1.49	Salt Bridge lost; key catalytic residue, very conserved	ss6694:10% (Multination) (Halushka et al. 1999)
rs5518	KLK1	NP_002248	V193E	-1.62	-0.70	Buried Charge, hydrophobic interaction decreased	ss6984:5% (Multination) (Halushka et al. 1999)

Table 4-5. Example candidate SNPs for hypertension

obstruction and widening and thickening of infrarenal arterial vessels' (Krege et al. 1995). The third SNP, H66R, is in chymase (CMA1), and changes a key catalytic residue, as well a breaking a salt bridge. The physiological function of chymase is still controversial (Ju et al. 2001) (Takai and Miyazaki 2003). A SNP upstream of the transcription initiation site of CMA1 has been reported to be associated with hypertensive complications such as HDL cholesterol (possibly related to its lipid metabolism function), but not with blood pressure (Fukuda et al. 2002). The fourth SNP, V193E, in kallikrein (KLK1) results in a buried charge and loss of hydrophobic burial, affecting bradykinin processing.

Discussion

There are three unique features of the SNPs3D resource. First, it is designed specifically for the analysis of the relationship between SNPs and disease. Second, it constructs gene networks based on conceptual relationships derived from the literature, rather than experimental data. Third, it integrates access to all available and relevant information sources, wherever possible giving the user easy access to the underlying data and literature, so that informed judgments can be made.

We have chosen to construct a network of connections between genes based on how strongly they are coupled in the literature, rather than whether there is extractable information supporting a physical interaction between them. There are two advantages to this approach. First, relevant connections between proteins may be non-physical. For example, genes that are involved in the same complex disease may not directly interact, or even be in the same local pathway, but may nevertheless interact in terms of affecting disease susceptibility. Second, the text mining procedure will capture considerably more information than is currently in any database, or that can be easily formalized in a simple cause and effect pathway description. In this sense, the KnowledgeNet expands on existing pathways descriptions by linking genes with conceptual relationships.

The case studies illustrate how all this works in practice. Analysis of non-synonymous SNPs in the selectins leads to the finding of several that appear to be deleterious to protein function, but which do not directly lead to a disease phenotype. Inspection of homologs in the KnowledgeNet graphical interface suggests a role for functional redundancy in conferring network level robustness, and consulting mouse knockout and expression profile data supports that conclusion. The result also strongly suggests an epistatic relationship between the deleterious SNPs in selectin E and selectin P: An individual homozygous in either one will likely not display clinical symptoms, but an individual homozygous in both will probably have a significantly compromised inflammatory response. In the hypertension example, a list of possible candidate genes is generated. The KnowledgeNet interface allows a user to browse the relationships between those genes, clustering the main pathways, and providing access to analysis of the relevant non-synonymous SNPs. As is often the case, the roles of the some of the genes in disease susceptibility are complicated, and the available information is some times contradictory. For example, for chymase, there is considerable uncertainty of function. Instant access to the relevant literature allows the user to quickly appreciate the subtleties of the current state of knowledge.

We now consider the strengths and weaknesses of the approach in more detail.

Concept profiles for genes are built from the relative frequency of words and terms in PubMed abstracts. In turn, overlap of the profiles are used to identify gene-gene relationships. In practice, the procedure provides intuitively reasonably results, but there is no way of rigorously benchmarking such knowledge generated networks. The method occasionally makes errors on the side of over-inclusiveness. For example, it is not able to distinguish between statements such as ‘protein A is associated with disease B’ versus ‘protein A is not associated with disease B’. As illustrated in the Results, it is also possible for a disease and gene to be linked by irrelevant factors, such as symptoms common to more than one syndrome. Similarly, gene-gene relationships may sometimes be based on non-pathway related factors. For example the 13 members of the human kallikrein family are tightly coupled, because of many articles that discuss them as a group. In fact, most of the family members operate in quite different pathways. In future, more sophisticated natural language processing technology may be applied to reduce these effects. At present, a concept overlap weighting scheme that emphasizes relationships to 'hub' proteins is used, and ensures that proteins weakly linked to these are included. A weighting scheme that takes into account the number of papers published on a gene may further improve inclusion of relevant weak links. The analysis is limited to abstracts already annotated as relevant to a particular gene. Extension to all pubmed abstracts (currently about 8.5 million) is desirable. In practice, the resource is very effective at narrowing down the amount of literature a user must consult in arriving at an informed position, our main goal.

Concept profile overlaps are also used to provide lists of candidate genes for involvement in susceptibility to particular diseases. There is no gold standard for candidate genes for a disease, with different compilations using different criteria. Comparison of our hypertension list with a hand compiled list for essential hypertension (Halushka et al. 1999), shows informative similarities and differences. That list contains 75 candidate genes rated as 'strong', 57 of which are also in the SNPs3D hypertension set. Nine of the top ten ranking SNPs3D genes are in the hand complied hypertension list. The exception is BMPR2, which is involved in pulmonary hypertension, rather than essential hypertension. The 12th ranking gene in the SNPs3D list, ADRB2, is also not in the hand complied list, but is clearly associated with hypertension in PubMed abstracts. Conversely, some of the additional genes in the hand complied list, such as GALR1, are not linked in any way to hypertension in PubMed, even with a more sophisticated profile based search, and including all abstracts. Their selection may reflect specialized insights on the part of the compliers. Others, such APOC2 and APOC4, are also not associated with hypertension in PubMed, but have a chromosome location covered by a known hypertension marker.

SNPs3D candidate lists can be generated on demand, with little delay, and so have the advantage of taking into account all the current literature. On the other hand, there is a great deal of relevant specialized knowledge in the scientific community that is either not in the literature, or very difficult to extract in a useful way. The Genetic Association Database (GAD) is an archive of human genetic association studies of complex diseases and disorders (Becker et al. 2004) that provides an alternative approach to compiling the relevant information. Any user may submit information about an association between a disease and a gene, creating a mechanism of capturing community knowledge. We expect that in the long run, the most effective candidate lists will be complied by a hybrid of the two approaches.

SNPs3D analysis is only provided for non-synonymous SNPs. Other sorts of SNPs, particularly those affecting transcription, splicing and perhaps RNA message structure will also play a role in susceptibility to complex trait disease. Little data on is available on the relative importance of the different SNP types, although for monogenic disease, the role is relatively small. For example, single base variant effects operating through transcription are quite rare, accounting for 0.5% of cases (Stenson et al. 2003). Whatever the case, it is clearly desirable to include other classes of SNP. It should shortly be possible to extend coverage in this way, using DNA sequence profiles based on the complete genome sequences of higher eukaryotes.

Methods

Query Interface

Each of the three modules (SNP analysis, gene-gene network, and disease candidate gene lists and networks) is accessed via a separate simple search window, on the site front page.

The candidate gene search window will accept any word or phrase as an entry, and compiles a concept profile, as described below. For SNP analysis and gene-gene networks requests, a hierarchal query string processing procedure is used, providing a wide choice of input name types, including dbSNP IDs, Entrez Gene IDs, RefSEQ IDs, NBCI Gene Symbols, and common protein names, using the following procedure:

1. A query string is first inspected to determine if its composition is consistent with a dbSNP ID, Entrez Gene ID or Refseq ID. If one of these name types is identified, the query is searched against the corresponding list of possibilities, and if a match is found, appropriate results are returned.

2. If the type of ID cannot be identified, the query string is first treated as a NCBI gene symbol, and searched against that set. If an exact match is found, results are returned.

3. If no exact match to a gene symbol is found, the string is searched against all words in the NCBI Gene summaries of each gene. Any hit adds to a list of high ranked possible genes.

4. This hit list is supplemented by a search of the query string against all the PubMed abstracts associated with each gene in the NCBI Gene Database. The number of times the query string is found in the abstracts for a gene provides a ranking weight. Finally, the user is invited to choose the appropriate gene from the ranked list of possibilities.

5. If a search completely fails, the user is offered an alternative search window, with explicit query string categories.

Literature Dataset

The abstracts of all the medline entries associated with each gene in the NCBI Gene database (Pruitt et al. 2000) are the source of words and terms. In the current version, there are, 80,249 Medline references linked to 19,228 human genes. Word types are identified using SVMtagger (http://www.lsi.upc.es/~nlp/SVMTool/). Keyterms are constructed from single nouns and adjectives, adjective/noun pairs, and continuous strings of words classified as adjectives or nouns. For example, the phrase ‘blood pressure’ occurring in an abstract would result in three keyterms: ‘blood’, ‘pressure’, and ‘blood pressure’. Terms occurring only once are removed. There are currently a total of 266,337 keyterms.

The number of occurrences of each keyterm ‘KW’ in all the abstracts (‘Total_count(KW)’ is retained, as well as the number of occurrences of each keyterm in the abstracts associated with each gene ‘G’, ‘Count(G,KW)’, and the fraction of all occurrences of each keyterm that are associated with each gene is calculated as:

F1(G,KW) = Count(G,KW)/Total_Count(KW)

Construction of the Gene-Gene Relationship Matrix

The interaction strength L(i,j) between every pair of genes i and j is calculated as:

$L(i,j) = \sum_{KW} F1(G_i,KW) + \sum_{KW} F1(G_j,KW)$

where the sum is over all keyterms common to the two genes, excluding any found in more than 300 genes. More studied genes have more associated abstracts in the NCBI Gene database, so that this expression upweights interactions involving those. Comparison with a more egalitarian gene-gene weighting, based on a dot product sum similar to that used for the disease/gene linkage, suggests that an emphasis on the hub-like genes is useful for including links to relevant but more weakly coupled genes.

Because of memory constraints, the interactions are stored as a sparse matrix, retaining a maximum of 200 interacting genes per gene. A few well studied genes, such as P53, have more than 200 genes linked with significant scores (greater than the mean element value of the sparse matrix). However, in almost all cases, these elements will be included in the list of associations for other genes.

Generation of a Candidate Gene List for a Disease

Given a disease name, a list of candidate genes is generated as follows:

A. The subset of abstracts relevant to the disease is identified:

1. Any abstract containing the full disease name, for example, ‘breast cancer’ is selected.

2. If this procedure results in less than 20 abstracts, and the disease name consists of more than one word, a further search of abstracts is made for the combination of words, for example ‘breast’ AND ‘cancer’.

3. If less than a total of ten abstracts are selected, the process is aborted, returning a message of ‘Not enough abstracts to build a profile’.

B: A keyterm profile is generated for the disease, using the selected abstracts. All Keyterms are ranked by the fraction of disease abstracts that contain them:

Rank(KW)= Count_abstracts(D,KW)/[Total_abstracts(KW) +50]

where ‘Count_abstracts(D,KW)’ is the number of abstracts for disease ‘D’ containing the keyterm ‘KW’, and ‘Total_abstracts(KW)’ is the total number of abstracts containing the keyterm. A pseudo count of 50 is added to reduce noise. The top ranking 40 keyterms are selected, providing Rank(KW) is at least 0.1.

C: The overlap of the disease keyterms with those of each gene is calculated:

1. The number of times each selected keyterm ‘KW’ occurs in the abstracts associated with the disease ‘D’, ‘Count(D,KW)’, is determined, and the relative frequency is calculated as :

F2(D,KW) = Count(D,KW)/Total_Count(KW)

2. The strength of association of the disease ‘D’ with a gene ‘G’ is calculated as the dot product of the relative frequencies of the disease keyterms with the relative frequencies of those same keyterms in that gene:

SD(D,G) = $\sum_{KW}$ F1(G,KW).F2(D,KW)

where the sum is only over the up to 40 keywords selected as the keyterm set of disease ‘D’. The association strength is deliberately biased towards the keyterms most strongly associated with the disease, as opposed to be associated with particular genes.

D: Finally, all genes with a non-zero score are returned as candidates.

Database Setup

The database is implemented in MySQL. As shown in figure 4-6, the central table is ‘Gene’, an up-to-date list of human genes from the NCBI Entrez Gene database. The Gene table is linked to other master tables: The SNP model table contains our stability and profile analysis of SNPs. There is a table of keyterms for each gene, and a table of PubMed abstract IDs for each gene. The KnowledgeNet matrix table contains the pairwise gene-gene interaction strengths, and there is also a disease/candidate gene matrix. Some other tables linked to the Gene table are: the Transcript table (RefSeq mRNAs); the Protein table (RefSeq proteins); the phenotype and disease-tables (NCBI OMIM and human gene mutation database (HGMD)); Mouse knockout table (Bioscience mouse knockout); pathway (KEGG), protein-protein interactions (BIND); and protein function (GO).

Web Interface

SNPs3D is served using Apache software running on a Linux PC and with web pages derived from an early open source version of PHP-NUKE (http://www.phpnuke.org/).

KnowledgeNet Graphical Interface

The interactive graphical interface for displaying gene-gene relationships is based on open source Java code (http://www.touchgraph.com). Genes form nodes in a

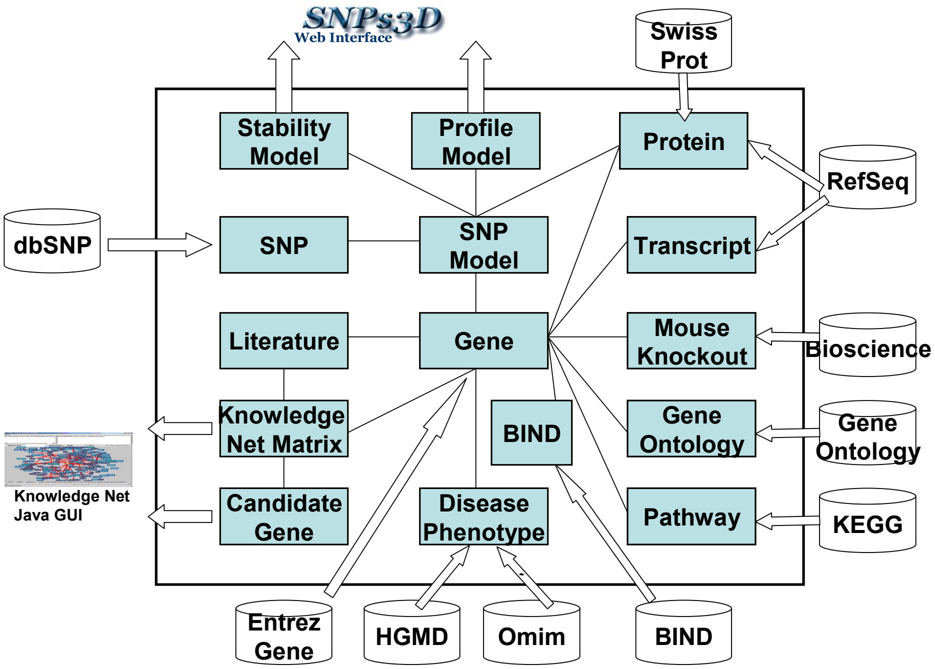


Figure 4-6. Database Schema.

The Blue blocks represent individual modules, which may be single or multiple MySql tables.

graph and gene-gene relationships are edges. Clicking links and symbols leads to more detailed information. Symbol shape; font style; symbol, edge and font color as well as hover-over windows are used to provide as much information as possible. Gene symbol shape conveys whether or not that gene is involved in disease, gene symbol text color indicates whether there are deleterious SNPs. Subsets of genes containing one or more SNPs with population frequencies above some threshold may be highlighted (identifying those most likely to be involved in complex traits). A maximum of 300 genes are displayed in the graphical interface. These are genes most strongly associated with a query gene or a query disease. The threshold for displaying links between genes is adjustable to show only those most strongly linked, or all possible connections. Links may also be based on KEGG pathway connections or direct protein-protein interaction information, extracted from BIND (Bader et al. 2003). Left clicking on a gene provides immediate access to all the gene specific information, including SNP analysis using the stability (Yue et al. 2005) and profile methods (Yue 2005) and the NCBI Gene summary, as well as pathways, dbSNP entries and homologs.

Content for the graphical display can be generated using the list of genes associated with a reference gene or a disease (the candidate genes, with the strongest linked gene as initial center), or a specified list of genes. All gene lists may be edited. One important feature is the ability to redraw the graph, using a selected node as the new center, allowing the user to smoothly navigate through adjacent regions of the knowledgeNet matrix. A pull down menu provides a list of all displayed genes, and any gene may be highlighted in the network via this list. Right clicking on a node provides facilities for highlighting genes which share certain properties with the reference gene, such as KEGG pathway, associated papers, or sequence homology. Left clicking in a gene brings up its SNP analysis.

Chapter 5: Discussion and Conclusion

Progress in Understanding Monogenic Disease

Role of Protein Destabilization in Monogenic Disease

This thesis describes an investigation into the mechanisms by which mis-sense variants (the most abundant known disease variants) cause human disease. Since *in vitro* mutagenesis data show that many single residue variations decrease protein stability, we hypothesized that loss of stability plays a major role in causing human monogenic diseases. In order to test this hypothesis, we developed a structure based model to evaluate the effect of mis-sense variants on protein stability by looking at 15 structure features, such as electrostatic interactions, and overpacking. The model successfully identifies 74% of mis-sense variants known to cause human monogenic disease with a 15% false positive rate. We therefore conclude that the majority of monogenic disease variants act by destabilization of protein structure.

Size of the Destabilization Effect

The stability model was applied to a set of destabilizing mutations for which the *in vitro* change in stability has been experimentally measured. We found that only a small fraction of mutants that stabilize or weakly destabilize a structure are assigned a disease-causing outcome, consistent with the overall false positive rate of the model, while 90% of mutants that destabilize a structure by 3 Kcal/mol or more are classified as disease causing. In addition to supporting the role of destabilization, this analysis provides an approximate free energy scale for disease-causing mutants – typically 2 to 3 Kcal/mol.

Significance of a Small Destabilization Effect in vivo

The free energy difference between the folded and unfolded state of a globular protein typically ranges from 5-15 Kcal/mol (Privalov 1979), corresponding to an equilibrium constant between 10^{-4} and 10^{-13}. A typical disease causing mutant destabilizes the folded state by 2 Kcal/mol and so increases the concentration of the unfolded state by about two orders of magnitude. However, the fraction of unfolded molecules is still very small, so such a mutant will usually not have a detectable effect *in vitro*. *In vivo,* the 100 fold increase in the concentration of the unfolded state will result in a proportional increase of scavenging by chaperones (Hohfeld et al. 2001). We propose that this mechanism may play a role in dramatically lowering the concentration of a disease mutant protein *in vivo*. However, further experiments are required to test this hypothesis.

Structure and Sequence Based Models: Pros and Cons

The structure based model allows us to investigate how an amino acid variant affects protein function. However, its application is limited by its requirement for protein three-dimensional (3D) structure. Experimentally determining human protein 3D structure is very challenging, and only a small fraction is so far available. Comparative modeling expands the useful structure coverage of human proteins, but still only 10% of human proteins can be analyzed by the stability model. The sequence model does not require protein structure information and thus has a broader application, and also detects a wider range of functional effects. It relies on analysis of evolutionary constraints which can be inferred from multiple alignments between human protein sequences and their homologs. The drawback of the sequence model, however, is that it can not provide direct insight on the mechanism by which an amino acid variant affects function. The primary errors in the two models are caused by different factors: an incorrect protein structure for the stability method and too few sequences in an alignment for the sequence method. The classification of a given variant can be further validated by comparing the results by these two models.

Analysis of Human Population SNPs

One-Fourth of Human Mis-sense SNPs are Deleterious

Both models were applied to known human population SNPs. One major conclusion of this thesis is that about one quarter of the known missense SNPs in the human population are significantly deleterious to protein function *in vivo*. Two factors have been carefully considered in reaching that conclusion. The first factor is related to the errors caused by the models and the second factor is related to the errors in the dataset. The rate of deleterious SNPs in the population is overestimated due to the false positives and underestimated because of the false negatives. We took both types of errors into account in estimation of deleterious SNPs. False positive and false negative rates were obtained from benchmarking the methods against monogenic disease and a fitting procedure was then used to find the true deleterious rate. The effect of errors in the dataset is controlled by comparing the results between all available SNPs and the SNPs validated by the HapMap project (2003) and Perlegen Inc (Hinds et al. 2005).

Discussion of Deleterious SNPs

Most of the newly identified deleterious SNPS are not in human monogenic disease genes, and do not have any known role in complex disease. There are three broad categories of explanation for this:

1. These SNPs are not deleterious to protein function as our models suggest. While this explanation can not be entirely ruled out, we have carefully taken into account the effect of errors in the models and in the data.

2. The monogenic disease proteins are somehow especially vulnerable to the effect of deleterious mis-sense mutants. For example, the mechanism of removing unfolded proteins by chaperone-dependent processes may be only applicable to monogenic disease genes. Little is known to prove or reject this possibility. A comparison of protein types between disease genes and non-disease genes, based on their GO classification of molecular function, does not reveal any significant difference.

3. The phenotype is somehow buffered against deleterious SNPs in most genes. That is, the decrease or loss of function of a single gene caused by a deleterious variant does not show any significant impact at the phenotype level. Such a hypothesis is supported by the results of knockout experiments. Gene suppression in C.elegans (Kamath et al. 2003) and Saccharomyces (Cliften et al. 2003) (Rubin et al. 2000) as well as limited mouse knockout data show that loss of function of many individual genes does not cause any detectable phenotype change. One possible buffering mechanism is redundant function between genes, especially among sequence paralogs (Kafri et al. 2005). Most human genes have paralogs within the

genome, but we found no difference between monogenic disease genes and non-disease genes. Therefore, functional redundancy between paralogous genes may not be the primary mechanism to buffer effects of damaging mutations *in vivo*. The second possible mechanism lies in the network properties of human gene-gene interactions. Many known pathways contain feedback loops and alternative routes that provide system level robustness against damaging mutations. Inspection of individual cases suggests that this type of buffering is the major factor.

The KnowledgeNet: a knowledge based gene-gene network

In order to understand the impact of a deleterious SNP on the phenotype, it is necessary to consider its network environment. We have constructed a knowledge-based gene-gene network using a simple text mining method. In the network, gene pairs are linked according to the overlap between their concept profiles. A concept profile is a simple means of capturing the concept associated with each gene in the literature. Each concept profile is a list of words and phrases found in abstracts related to a gene. The advantage of such a network lies in its inclusiveness, because it reflects not only the known physical interactions between different genes, but also more abstract relationships between them. For example, two genes may be linked because they are both involved in the same disease even though they may not directly interact. The disadvantage of this network is also apparent: precise definitions of the relationship between gene pairs are not available.

Concept profiles are not limited to genes, but can also be compiled for diseases or biological processes. Overlap of disease and gene profiles can be used to compile a list of candidate genes involved in a given disease. The case study of hypertension shows that such an automatically-generated candidate list does include most genes in known blood pressure regulation related pathways. It also suggests some new genes that are missing from the available expert curated gene list (Halushka et al. 1999).

Relevance to Public Health

Up to now, most attempts to link SNPs and susceptibility to complex disease have relied on statistical association of a SNP with a disease. While there have been some noticeable successes, for example the role of APOE SNPs in Alzhemer's disease, in general it has proven difficult to relate mutations to disease. There are several possible explanations for this, including the role of epistasis effects (non-linear interactions between SNPs), the small contribution of most SNPs to disease susceptibility, and not including the relevant regions of the genome. Whole genome association studies are now being proposed to address the third of these possibilities. It is not yet known how effective the studies will be, and new problems of statistical significance are raised.

Understanding the mechanisms by which SNPs are related to disease offers a different and complementary approach to identifying disease mutations. The work described in this thesis covers one aspect of mechanism, and has several direct applications: 1) potentially deleterious human population SNPs are identified and thus provides a list of SNPs for association studies. 2) An automatically generated candidate gene list for a given disease can help an investigator in designing a candidate gene based association study 3) The gene-gene interaction network can help users investigate possible epistasis effects between candidate genes. More importantly, this work can contribute to understanding of mechanism of common human diseases by helping to address the following questions: How does a mutation affect protein function? How is the effect on protein function transformed into an effect at a phenotypic level? How do the network properties of gene-gene interactions buffer the effect of damage to a single gene?

Summary of Conclusions and Contributions

The major conclusions and contributions of this thesis can be summarized as follows:

We conclude that the loss of stability plays a major role in the development of monogenic human diseases.

We conclude that approximately 25% of mis-sense SNPs in the human population significantly damage protein function. These mis-sense SNPs provide a list of candidates for association with common human diseases.

A simple gene-gene relationship network is set up to facilitate identification of network properties. The network allows investigation of the impact of mis-sense SNPs on phenotypes and identification of sets of mis-sense SNPs for incorporating epitasis effect into general association studies.

Concept profiles provide a means to identify links between gene and disease, allowing candidate genes to be compiled.

A website has been developed to allow free access to the data for the scientific community.

Limitations and Suggestions for Future Work

Other Mutations Affecting Disease Susceptibility:

There are many ways in which a SNP may affect a human phenotype. This work focuses only on the study of mis-sense SNPs because they are the most abundant genetic mutations causing human monogenic diseases. However, SNPs in gene regulatory regions have long been suspected to play a major role in common human disease. In addition, there are many non-gene related regions of DNA that display a high level of conservation between species of higher Eukaryotes, suggesting unknown but important functions (Loots et al. 2000).

The sequences of a number of higher Eukaryote genomes, including human, mouse, rat and chimpanzee, have been completed and more sequencing efforts are ongoing. With these data available, the principle of the amino acid sequence conservation model can be applied to analysis of genome conservation at the DNA level, identifying other classes of deleterious SNPs. Moreover, systematic experimental projects, such as ENCODE (2004), will also expand our knowledge of the function of these non-coding regions.

Beside SNPs, genomic structure variations, such as insertion, deletion and chromosomal duplication, have been observed in many cases of monogenic disease. As to common human diseases, the role of chromosome duplication in cancer has been broadly investigated, but so far only seldom studied in other diseases. Compared to SNPs, genomic structure variants are not easily detected. In future, new technologies may provide the necessary data and thus allow investigation into the wide genomic structure of human disease.

Gene-Gene Network Construction

The current gene-gene network is based on the over eighty thousand gene-related Medline abstracts in the NCBI Gene database. A general method of automatically identifying papers related to particular genes will broaden the coverage of the network. The current simple literature mining method also has its limitations. For example, a paper may state that protein A is not associated with disease C. The KnowledgeNet will ignore the ‘not’ and simply extract disease C as one of the keywords associated with protein A and thus erroneously link A to those proteins that are truly related to disease C. In future, natural language processing technology should be able to reduce these problems.

Experimental genetic approaches have been used in model organisms to systematically identify gene-gene interaction properties inside biological networks. A recent paper describes a system-level study on epistasis by single and double knockout of 890 metabolic genes in yeast (Segre et al. 2005). Incorporating these types of new data into the KnowledgeNet will further increase its usefulness.

References

2003. The International HapMap Project. *Nature* **426:** 789-796.

2004. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* **306:** 636-640.

Alfarano, C., C.E. Andrade, K. Anthony, N. Bahroos, M. Bajec, K. Bantoft, D. Betel, B. Bobechko, K. Boutilier, E. Burgess, K. Buzadzija, R. Cavero, C. D'Abreo, I. Donaldson, D. Dorairajoo, M.J. Dumontier, M.R. Dumontier, V. Earles, R. Farrall, H. Feldman, E. Garderman, Y. Gong, R. Gonzaga, V. Grytsan, E. Gryz, V. Gu, E. Haldorsen, A. Halupa, R. Haw, A. Hrvojic, L. Hurrell, R. Isserlin, F. Jack, F. Juma, A. Khan, T. Kon, S. Konopinsky, V. Le, E. Lee, S. Ling, M. Magidin, J. Moniakis, J. Montojo, S. Moore, B. Muskat, I. Ng, J.P. Paraiso, B. Parker, G. Pintilie, R. Pirone, J.J. Salama, S. Sgro, T. Shan, Y. Shu, J. Siew, D. Skinner, K. Snyder, R. Stasiuk, D. Strumpf, B. Tuekam, S. Tao, Z. Wang, M. White, R. Willis, C. Wolting, S. Wong, A. Wrong, C. Xin, R. Yao, B. Yates, S. Zhang, K. Zheng, T. Pawson, B.F. Ouellette, and C.W. Hogue. 2005. The Biomolecular Interaction Network Database and related tools 2005 update. *Nucleic Acids Res* **33:** D418-424.

Altschul, S.F., T.L. Madden, A.A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D.J. Lipman. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25:** 3389-3402.

Altshuler, D., J.N. Hirschhorn, M. Klannemark, C.M. Lindgren, M.C. Vohl, J. Nemesh, C.R. Lane, S.F. Schaffner, S. Bolk, C. Brewer, T. Tuomi, D. Gaudet, T.J. Hudson, M. Daly, L. Groop, and E.S. Lander. 2000. The common PPARgamma Pro12Ala polymorphism is associated with decreased risk of type 2 diabetes. *Nat Genet* **26:** 76-80.

Andreeva, A., D. Howorth, S.E. Brenner, T.J. Hubbard, C. Chothia, and A.G. Murzin. 2004. SCOP database in 2004: refinements integrate structure and sequence family data. *Nucleic Acids Res* **32:** D226-229.

Bader, G.D., D. Betel, and C.W. Hogue. 2003. BIND: the Biomolecular Interaction Network Database. *Nucleic Acids Res* **31:** 248-250.

Baker, D. and A. Sali. 2001. Protein structure prediction and structural genomics. *Science* **294:** 93-96.

Bava, K.A., M.M. Gromiha, H. Uedaira, K. Kitajima, and A. Sarai. 2004. ProTherm, version 4.0: thermodynamic database for proteins and mutants. *Nucleic Acids Res* **32:** D120-121.

Becker, K.G., K.C. Barnes, T.J. Bright, and S.A. Wang. 2004. The genetic association database. *Nat Genet* **36:** 431-432.

Beveridge, D.L. and F.M. DiCapua. 1989. Free energy via molecular simulation: applications to chemical and biomolecular systems. *Annu Rev Biophys Biophys Chem* **18:** 431-492.

Bhasin, M., H. Zhang, E.L. Reinherz, and P.A. Reche. 2005. Prediction of methylated CpGs in DNA sequences using a support vector machine. *FEBS Lett* **579:** 4302-4308.

Bidiwala, S. and T. Pittman. 2004. Neural network classification of pediatric posterior fossa tumors using clinical and imaging data. *Pediatr Neurosurg* **40:** 8-15.

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

Botstein, D. and N. Risch. 2003. Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nat Genet* **33 Suppl:** 228-237.

Byrne, M.P., R.L. Manuel, L.G. Lowe, and W.E. Stites. 1995. Energetic contribution of side chain hydrogen bonding to the stability of staphylococcal nuclease. *Biochemistry* **34:** 13949-13960.

Canutescu, A.A., A.A. Shelenkov, and R.L. Dunbrack, Jr. 2003. A graph-theory algorithm for rapid protein side-chain prediction. *Protein Sci* **12:** 2001-2014.

Cargill, M., D. Altshuler, J. Ireland, P. Sklar, K. Ardlie, N. Patil, N. Shaw, C.R. Lane, E.P. Lim, N. Kalyanaraman, J. Nemesh, L. Ziaugra, L. Friedland, A. Rolfe, J. Warrington, R. Lipshutz, G.Q. Daley, and E.S. Lander. 1999. Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nat Genet* **22:** 231-238.

Carlson, C.S., M.A. Eberle, L. Kruglyak, and D.A. Nickerson. 2004. Mapping complex disease loci in whole-genome association studies. *Nature* **429:** 446-452.

Cavallo, A. and A.C. Martin. 2005. Mapping SNPs to protein sequence and structure data. *Bioinformatics* **21:** 1443-1450.

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

Cliften, P., P. Sudarsanam, A. Desikan, L. Fulton, B. Fulton, J. Majors, R. Waterston, B.A. Cohen, and M. Johnston. 2003. Finding functional features in Saccharomyces genomes by phylogenetic footprinting. *Science* **301:** 71-76.

Corder, E.H., A.M. Saunders, W.J. Strittmatter, D.E. Schmechel, P.C. Gaskell, G.W. Small, A.D. Roses, J.L. Haines, and M.A. Pericak-Vance. 1993. Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families. *Science* **261:** 921-923.

Cozzetto, D. and A. Tramontano. 2005. Relationship between multiple sequence alignments and quality of protein comparative models. *Proteins* **58:** 151-157.

Dantzer, J., C. Moad, R. Heiland, and S. Mooney. 2005. MutDB services: interactive structural analysis of mutation data. *Nucleic Acids Res* **33:** W311-314.

Daraselia, N., A. Yuryev, S. Egorov, S. Novichkova, A. Nikitin, and I. Mazo. 2004. Extracting human protein interactions from MEDLINE using a full-sentence parser. *Bioinformatics* **20:** 604-611.

Deshpande, N., K.J. Addess, W.F. Bluhm, J.C. Merino-Ott, W. Townsend-Merino, Q. Zhang, C. Knezevich, L. Xie, L. Chen, Z. Feng, R.K. Green, J.L. Flippen-Anderson, J. Westbrook, H.M. Berman, and P.E. Bourne. 2005. The RCSB Protein Data Bank: a redesigned query system and relational database based on the mmCIF schema. *Nucleic Acids Res* **33 Database Issue:** D233-237.

Dodge, C., R. Schneider, and C. Sander. 1998. The HSSP database of protein structure-sequence alignments and family profiles. *Nucleic Acids Res* **26:** 313-315.

Emahazion, T., L. Feuk, M. Jobs, S.L. Sawyer, D. Fredman, D. St Clair, J.A. Prince, and A.J. Brookes. 2001. SNP association studies in Alzheimer's disease highlight problems for complex disease analysis. *Trends Genet* **17:** 407-413.

Ferrer-Costa, C., M. Orozco, and X. de la Cruz. 2002. Characterization of disease-associated single amino acid polymorphisms in terms of sequence and structure properties. *J Mol Biol* **315:** 771-786.

Fields, S. and O. Song. 1989. A novel genetic system to detect protein-protein interactions. *Nature* **340:** 245-246.

Frenette, P.S., T.N. Mayadas, H. Rayburn, R.O. Hynes, and D.D. Wagner. 1996. Susceptibility to infection and altered hematopoiesis in mice deficient in both P- and E-selectins. *Cell* **84:** 563-574.

Fukuda, M., T. Ohkubo, T. Katsuya, A. Hozawa, T. Asai, M. Matsubara, H. Kitaoka, I. Tsuji, T. Araki, H. Satoh, J. Higaki, S. Hisamichi, Y. Imai, and T. Ogihara. 2002. Association of a mast cell chymase gene variant with HDL cholesterol, but not with blood pressure in the Ohasama study. *Hypertens Res* **25:** 179-184.

Fulton, K.F., E.R. Main, V. Daggett, and S.E. Jackson. 1999. Mapping the interactions present in the transition state for unfolding/folding of FKBP12. *J Mol Biol* **291:** 445-461.

Gabriel, S.B., S.F. Schaffner, H. Nguyen, J.M. Moore, J. Roy, B. Blumenstiel, J. Higgins, M. DeFelice, A. Lochner, M. Faggart, S.N. Liu-Cordero, C. Rotimi, A. Adeyemo, R. Cooper, R. Ward, E.S. Lander, M.J. Daly, and D. Altshuler. 2002. The structure of haplotype blocks in the human genome. *Science* **296:** 2225-2229.

Giot, L., J.S. Bader, C. Brouwer, A. Chaudhuri, B. Kuang, Y. Li, Y.L. Hao, C.E. Ooi, B. Godwin, E. Vitols, G. Vijayadamodar, P. Pochart, H. Machineni, M. Welsh, Y. Kong, B. Zerhusen, R. Malcolm, Z. Varrone, A. Collis, M. Minto, S. Burgess, L. McDaniel, E. Stimpson, F. Spriggs, J. Williams, K. Neurath, N. Ioime, M. Agee, E. Voss, K. Furtak, R. Renzulli, N. Aanensen, S. Carrolla, E. Bickelhaupt, Y. Lazovatsky, A. DaSilva, J. Zhong, C.A. Stanyon, R.L. Finley, Jr., K.P. White, M. Braverman, T. Jarvie, S. Gold, M. Leach, J. Knight, R.A. Shimkets, M.P. McKenna, J. Chant, and J.M. Rothberg. 2003. A protein interaction map of Drosophila melanogaster. *Science* **302:** 1727-1736.

Green, S.M., A.K. Meeker, and D. Shortle. 1992. Contributions of the polar, uncharged amino acids to the stability of staphylococcal nuclease: evidence for mutational effects on the free energy of the denatured state. *Biochemistry* **31:** 5717-5728.

Green, S.M. and D. Shortle. 1993. Patterns of nonadditivity between pairs of stability mutations in staphylococcal nuclease. *Biochemistry* **32:** 10131-10139.

Gromiha, M.M., S. Ahmad, and M. Suwa. 2004. Neural network-based prediction of transmembrane beta-strand segments in outer membrane proteins. *J Comput Chem* **25:** 762-767.

Guerois, R., J.E. Nielsen, and L. Serrano. 2002. Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. *J Mol Biol* **320:** 369-387.

Halushka, M.K., J.B. Fan, K. Bentley, L. Hsie, N. Shen, A. Weder, R. Cooper, R. Lipshutz, and A. Chakravarti. 1999. Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis. *Nat Genet* **22:** 239-247.

Hamosh, A., A.F. Scott, J.S. Amberger, C.A. Bocchini, and V.A. McKusick. 2005. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res* **33:** D514-517.

Harris, M.A., J. Clark, A. Ireland, J. Lomax, M. Ashburner, R. Foulger, K. Eilbeck, S. Lewis, B. Marshall, C. Mungall, J. Richter, G.M. Rubin, J.A. Blake, C. Bult, M. Dolan, H. Drabkin, J.T. Eppig, D.P. Hill, L. Ni, M. Ringwald, R. Balakrishnan, J.M. Cherry, K.R. Christie, M.C. Costanzo, S.S. Dwight, S. Engel, D.G. Fisk, J.E. Hirschman, E.L. Hong, R.S. Nash, A. Sethuraman, C.L. Theesfeld, D. Botstein, K. Dolinski, B. Feierbach, T. Berardini, S. Mundodi, S.Y. Rhee, R. Apweiler, D. Barrell, E. Camon, E. Dimmer, V. Lee, R. Chisholm, P. Gaudet, W. Kibbe, R. Kishore, E.M. Schwarz, P. Sternberg, M. Gwinn, L. Hannick, J. Wortman, M. Berriman, V. Wood, N. de la Cruz, P. Tonellato, P. Jaiswal, T. Seigfried, and R. White. 2004. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res* **32 Database issue:** D258-261.

Henikoff, S. and J.G. Henikoff. 1993. Performance evaluation of amino acid substitution matrices. *Proteins* **17:** 49-61.

Henrick, K. and J.M. Thornton. 1998. PQS: a protein quaternary structure file server. *Trends Biochem Sci* **23:** 358-361.

Hinds, D.A., L.L. Stuve, G.B. Nilsen, E. Halperin, E. Eskin, D.G. Ballinger, K.A. Frazer, and D.R. Cox. 2005. Whole-genome patterns of common DNA variation in three human populations. *Science* **307:** 1072-1079.

Hirschhorn, J.N. and M.J. Daly. 2005. Genome-wide association studies for common diseases and complex traits. *Nat Rev Genet* **6:** 95-108.

Hohfeld, J., D.M. Cyr, and C. Patterson. 2001. From the cradle to the grave: molecular chaperones that may choose between folding and degradation. *EMBO Rep* **2:** 885-890.

Holder, J.B., A.F. Bennett, J. Chen, D.S. Spencer, M.P. Byrne, and W.E. Stites. 2001. Energetics of side chain packing in staphylococcal nuclease assessed by exchange of valines, isoleucines, and leucines. *Biochemistry* **40:** 13998-14003.

Hua, S. and Z. Sun. 2001. A novel method of protein secondary structure prediction with high segment overlap measure: support vector machine approach. *J Mol Biol* **308:** 397-407.

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

Jansen, R., H. Yu, D. Greenbaum, Y. Kluger, N.J. Krogan, S. Chung, A. Emili, M. Snyder, J.F. Greenblatt, and M. Gerstein. 2003. A Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science* **302:** 449-453.

Ju, H., R. Gros, X. You, S. Tsang, M. Husain, and M. Rabinovitch. 2001. Conditional and targeted overexpression of vascular chymase causes hypertension in transgenic mice. *Proc Natl Acad Sci U S A* **98:** 7469-7474.

Kafri, R., A. Bar-Even, and Y. Pilpel. 2005. Transcription control reprogramming in genetic backup circuits. *Nat Genet* **37:** 295-299.

Kamath, R.S., A.G. Fraser, Y. Dong, G. Poulin, R. Durbin, M. Gotta, A. Kanapin, N. Le Bot, S. Moreno, M. Sohrmann, D.P. Welchman, P. Zipperlen, and J. Ahringer. 2003. Systematic functional analysis of the Caenorhabditis elegans genome using RNAi. *Nature* **421:** 231-237.

Kanehisa, M., S. Goto, S. Kawashima, Y. Okuno, and M. Hattori. 2004. The KEGG resource for deciphering the genome. *Nucleic Acids Res* **32:** D277-280.

Kragelund, B.B., P. Osmark, T.B. Neergaard, J. Schiodt, K. Kristiansen, J. Knudsen, and F.M. Poulsen. 1999. The formation of a native-like structure containing eight conserved hydrophobic residues is rate limiting in two-state protein folding of ACBP. *Nat Struct Biol* **6:** 594-601.

Krege, J.H., S.W. John, L.L. Langenbach, J.B. Hodgin, J.R. Hagaman, E.S. Bachman, J.C. Jennette, D.A. O'Brien, and O. Smithies. 1995. Male-female differences in fertility and blood pressure in ACE-deficient mice. *Nature* **375:** 146-148.

Krishnan, V.G. and D.R. Westhead. 2003. A comparative study of machine-learning methods to predict the effects of single nucleotide polymorphisms on protein function. *Bioinformatics* **19:** 2199-2209.

Kruglyak, L. and D.A. Nickerson. 2001. Variation is the spice of life. *Nat Genet* **27:** 234-236.

Lander, E.S. L.M. Linton B. Birren C. Nusbaum M.C. Zody J. Baldwin K. Devon K. Dewar M. Doyle W. FitzHugh R. Funke D. Gage K. Harris A. Heaford J. Howland L. Kann J. Lehoczky R. LeVine P. McEwan K. McKernan J. Meldrim J.P. Mesirov C. Miranda W. Morris J. Naylor C. Raymond M. Rosetti R. Santos A. Sheridan C. Sougnez N. Stange-Thomann N. Stojanovic A. Subramanian D. Wyman J. Rogers J. Sulston R. Ainscough S. Beck D. Bentley J. Burton C. Clee N. Carter A. Coulson R. Deadman P. Deloukas A. Dunham I. Dunham R. Durbin L. French D. Grafham S. Gregory T. Hubbard S. Humphray A. Hunt M. Jones C. Lloyd A. McMurray L. Matthews S. Mercer S. Milne J.C. Mullikin A. Mungall R. Plumb M. Ross R. Shownkeen S. Sims R.H. Waterston R.K. Wilson L.W. Hillier J.D. McPherson M.A. Marra E.R. Mardis L.A. Fulton A.T. Chinwalla K.H. Pepin W.R. Gish S.L. Chissoe M.C. Wendl K.D. Delehaunty T.L. Miner A. Delehaunty J.B. Kramer L.L. Cook R.S. Fulton D.L. Johnson P.J. Minx S.W. Clifton T. Hawkins E. Branscomb P. Predki P. Richardson S. Wenning T. Slezak N. Doggett J.F. Cheng A. Olsen S. Lucas C. Elkin E. Uberbacher M. Frazier R.A. Gibbs D.M. Muzny S.E. Scherer J.B. Bouck E.J. Sodergren K.C. Worley C.M. Rives J.H. Gorrell M.L. Metzker S.L. Naylor R.S. Kucherlapati D.L. Nelson G.M.

Weinstock Y. Sakaki A. Fujiyama M. Hattori T. Yada A. Toyoda T. Itoh C. Kawagoe H. Watanabe Y. Totoki T. Taylor J. Weissenbach R. Heilig W. Saurin F. Artiguenave P. Brottier T. Bruls E. Pelletier C. Robert P. Wincker D.R. Smith L. Doucette-Stamm M. Rubenfield K. Weinstock H.M. Lee J. Dubois A. Rosenthal M. Platzer G. Nyakatura S. Taudien A. Rump H. Yang J. Yu J. Wang G. Huang J. Gu L. Hood L. Rowen A. Madan S. Qin R.W. Davis N.A. Federspiel A.P. Abola M.J. Proctor R.M. Myers J. Schmutz M. Dickson J. Grimwood D.R. Cox M.V. Olson R. Kaul N. Shimizu K. Kawasaki S. Minoshima G.A. Evans M. Athanasiou R. Schultz B.A. Roe F. Chen H. Pan J. Ramser H. Lehrach R. Reinhardt W.R. McCombie M. de la Bastide N. Dedhia H. Blocker K. Hornischer G. Nordsiek R. Agarwala L. Aravind J.A. Bailey A. Bateman S. Batzoglou E. Birney P. Bork D.G. Brown C.B. Burge L. Cerutti H.C. Chen D. Church M. Clamp R.R. Copley T. Doerks S.R. Eddy E.E. Eichler T.S. Furey J. Galagan J.G. Gilbert C. Harmon Y. Hayashizaki D. Haussler H. Hermjakob K. Hokamp W. Jang L.S. Johnson T.A. Jones S. Kasif A. Kaspryzk S. Kennedy W.J. Kent P. Kitts E.V. Koonin I. Korf D. Kulp D. Lancet T.M. Lowe A. McLysaght T. Mikkelsen J.V. Moran N. Mulder V.J. Pollara C.P. Ponting G. Schuler J. Schultz G. Slater A.F. Smit E. Stupka J. Szustakowski D. Thierry-Mieg J. Thierry-Mieg L. Wagner J. Wallis R. Wheeler A. Williams Y.I. Wolf K.H. Wolfe S.P. Yang R.F. Yeh F. Collins M.S. Guyer J. Peterson A. Felsenfeld K.A. Wetterstrand A. Patrinos M.J. Morgan P. de Jong J.J. Catanese K. Osoegawa H. Shizuya S. Choi and Y.J. Chen. 2001. Initial sequencing and analysis of the human genome. *Nature* **409:** 860-921.

Laskowski RA, M.M., Moss DS, Thornton JM. 1993. PROCHECK: A program to check the stereochemical quality of protein structures. *J. Appl. Cryst.* **26:** 283-291.

Lau, A.Y. and D.I. Chasman. 2004. Functional classification of proteins and protein variants. *Proc Natl Acad Sci U S A* **101:** 6576-6581.

Lee, B. and F.M. Richards. 1971. The interpretation of protein structures: estimation of static accessibility. *J Mol Biol* **55:** 379-400.

Lee, I., S.V. Date, A.T. Adai, and E.M. Marcotte. 2004. A probabilistic functional network of yeast genes. *Science* **306:** 1555-1558.

Li, S., C.M. Armstrong, N. Bertin, H. Ge, S. Milstein, M. Boxem, P.O. Vidalain, J.D. Han, A. Chesneau, T. Hao, D.S. Goldberg, N. Li, M. Martinez, J.F. Rual, P. Lamesch, L. Xu, M. Tewari, S.L. Wong, L.V. Zhang, G.F. Berriz, L. Jacotot, P. Vaglio, J. Reboul, T. Hirozane-Kishikawa, Q. Li, H.W. Gabel, A. Elewa, B. Baumgartner, D.J. Rose, H. Yu, S. Bosak, R. Sequerra, A. Fraser, S.E.

Mango, W.M. Saxton, S. Strome, S. Van Den Heuvel, F. Piano, J. Vandenhaute, C. Sardet, M. Gerstein, L. Doucette-Stamm, K.C. Gunsalus, J.W. Harper, M.E. Cusick, F.P. Roth, D.E. Hill, and M. Vidal. 2004. A map of the interactome network of the metazoan C. elegans. *Science* **303:** 540-543.

Li, Z. and C. Chan. 2004. Inferring pathways and networks with a Bayesian framework. *Faseb J.*

Lifton, R.P., A.G. Gharavi, and D.S. Geller. 2001. Molecular mechanisms of human hypertension. *Cell* **104:** 545-556.

Loots, G.G., R.M. Locksley, C.M. Blankespoor, Z.E. Wang, W. Miller, E.M. Rubin, and K.A. Frazer. 2000. Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons. *Science* **288:** 136-140.

Main, E.R., K.F. Fulton, and S.E. Jackson. 1998. Context-dependent nature of destabilizing mutations on the stability of FKBP12. *Biochemistry* **37:** 6145-6153.

Mark, A.E. and W.F. van Gunsteren. 1994. Decomposition of the free energy of a system in terms of specific interactions. Implications for theoretical and experimental studies. *J Mol Biol* **240:** 167-176.

Markey, M.K., G.D. Tourassi, and C.E. Floyd, Jr. 2003. Decision tree classification of proteins identified by mass spectrometry of blood serum samples from people with and without lung cancer. *Proteomics* **3:** 1678-1679.

Markiewicz, P., L.G. Kleina, C. Cruz, S. Ehret, and J.H. Miller. 1994. Genetic studies of the lac repressor. XIV. Analysis of 4000 altered Escherichia coli lac repressors reveals essential and non-essential residues, as well as "spacers" which do not require a specific sequence. *J Mol Biol* **240:** 421-433.

Meeker, A.K., B. Garcia-Moreno, and D. Shortle. 1996. Contributions of the ionizable amino acids to the stability of staphylococcal nuclease. *Biochemistry* **35:** 6443-6449.

Mitsumori, T., S. Fation, M. Murata, K. Doi, and H. Doi. 2005. Gene/protein name recognition based on support vector machine using dictionary as features. *BMC Bioinformatics* **6 Suppl 1:** S8.

Molchanova, T.P., D.D. Pobedimskaya, and V. Postnikov Yu. 1994. A simplified procedure for sequencing amplified DNA containing the alpha 2- or alpha 1-globin gene. *Hemoglobin* **18:** 251-255.

Mooney, S.D. and R.B. Altman. 2003. MutDB: annotating human variation with functionally relevant data. *Bioinformatics* **19:** 1858-1860.

Mori, Y., K. Takeda, M. Charbonneau, and S. Refetoff. 1990. Replacement of Leu227 by Pro in thyroxine-binding globulin (TBG) is associated with complete TBG deficiency in three of eight families with this inherited defect. *J Clin Endocrinol Metab* **70:** 804-809.

Nariai, N., S. Kim, S. Imoto, and S. Miyano. 2004. Using protein-protein interactions for refining gene networks estimated from microarray data by Bayesian networks. *Pac Symp Biocomput***:** 336-347.

Ng, P.C., J.G. Henikoff, and S. Henikoff. 2000. PHAT: a transmembrane-specific substitution matrix. Predicted hydrophobic and transmembrane. *Bioinformatics* **16:** 760-766.

Ng, P.C. and S. Henikoff. 2003. SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res* **31:** 3812-3814.

O'Donnell, C.J., K. Lindpaintner, M.G. Larson, V.S. Rao, J.M. Ordovas, E.J. Schaefer, R.H. Myers, and D. Levy. 1998. Evidence for association and genetic linkage of the angiotensin-converting enzyme locus with hypertension and blood pressure in men but not women in the Framingham Heart Study. *Circulation* **97:** 1766-1772.

Oliva, M.T. and J. Moult. 1999. Local electrostatic optimization in proteins. *Protein Eng* **12:** 727-735.

Pan, Y. and V. Daggett. 2001. Direct comparison of experimental and calculated folding free energies for hydrophobic deletion mutants of chymotrypsin inhibitor 2: free energy perturbation calculations using transition and denatured states from molecular dynamics simulations of unfolding. *Biochemistry* **40:** 2723-2731.

Pedersen, J.T. and J. Moult. 1997. Protein folding simulations with genetic algorithms and a detailed molecular description. *J Mol Biol* **269:** 240-259.

Pelletier, J. and N. Sonenberg. 1987. The involvement of mRNA secondary structure in protein synthesis. *Biochem Cell Biol* **65:** 576-581.

Peri, S., J.D. Navarro, T.Z. Kristiansen, R. Amanchy, V. Surendranath, B. Muthusamy, T.K. Gandhi, K.N. Chandrika, N. Deshpande, S. Suresh, B.P. Rashmi, K. Shanker, N. Padma, V. Niranjan, H.C. Harsha, N. Talreja, B.M. Vrushabendra, M.A. Ramya, A.J. Yatish, M. Joy, H.N. Shivashankar, M.P. Kavitha, M. Menezes, D.R. Choudhury, N. Ghosh, R. Saravana, S. Chandran, S. Mohan, C.K. Jonnalagadda, C.K. Prasad, C. Kumar-Sinha, K.S. Deshpande, and A. Pandey. 2004. Human protein reference database as a discovery resource for proteomics. *Nucleic Acids Res* **32:** D497-501.

Phizicky, E., P.I. Bastiaens, H. Zhu, M. Snyder, and S. Fields. 2003. Protein analysis on a proteomic scale. *Nature* **422:** 208-215.

Plemper, R.K. and D.H. Wolf. 1999. Retrograde protein translocation: ERADication of secretory proteins in health and disease. *Trends Biochem Sci* **24:** 266-270.

Prince, J.A., L. Feuk, S.L. Sawyer, J. Gottfries, A. Ricksten, K. Nagga, N. Bogdanovic, K. Blennow, and A.J. Brookes. 2001. Lack of replication of association findings in complex disease: an analysis of 15 polymorphisms in prior candidate genes for sporadic Alzheimer's disease. *Eur J Hum Genet* **9:** 437-444.

Pritchard, J.K. and N.J. Cox. 2002. The allelic architecture of human disease genes: common disease-common variant...or not? *Hum Mol Genet* **11:** 2417-2423.

Pritchard, J.K. and M. Przeworski. 2001. Linkage disequilibrium in humans: models and data. *Am J Hum Genet* **69:** 1-14.

Privalov, P.L. 1979. Stability of proteins: small globular proteins. *Adv Protein Chem* **33:** 167-241.

Pruitt, K.D., K.S. Katz, H. Sicotte, and D.R. Maglott. 2000. Introducing RefSeq and LocusLink: curated human genome resources at the NCBI. *Trends Genet* **16:** 44-47.

Przybylski, D. and B. Rost. 2002. Alignments grow, secondary structure prediction improves. *Proteins* **46:** 197-205.

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

Reich, D.E. and E.S. Lander. 2001. On the allelic spectrum of human disease. *Trends Genet* **17:** 502-510.

Rennell, D., S.E. Bouvier, L.W. Hardy, and A.R. Poteete. 1991. Systematic mutation of bacteriophage T4 lysozyme. *J Mol Biol* **222:** 67-88.

Reumers, J., J. Schymkowitz, J. Ferkinghoff-Borg, F. Stricher, L. Serrano, and F. Rousseau. 2005. SNPeffect: a database mapping molecular phenotypic effects of human non-synonymous coding SNPs. *Nucleic Acids Res* **33:** D527-532.

Risch, N. and K. Merikangas. 1996. The future of genetic studies of complex human diseases. *Science* **273:** 1516-1517.

Rost, B. and V.A. Eyrich. 2001. EVA: large-scale analysis of secondary structure prediction. *Proteins* **Suppl 5:** 192-199.

Rubin, G.M., M.D. Yandell, J.R. Wortman, G.L. Gabor Miklos, C.R. Nelson, I.K. Hariharan, M.E. Fortini, P.W. Li, R. Apweiler, W. Fleischmann, J.M. Cherry, S. Henikoff, M.P. Skupski, S. Misra, M. Ashburner, E. Birney, M.S. Boguski, T. Brody, P. Brokstein, S.E. Celniker, S.A. Chervitz, D. Coates, A. Cravchik, A. Gabrielian, R.F. Galle, W.M. Gelbart, R.A. George, L.S. Goldstein, F. Gong, P. Guan, N.L. Harris, B.A. Hay, R.A. Hoskins, J. Li, Z. Li, R.O. Hynes, S.J. Jones, P.M. Kuehl, B. Lemaitre, J.T. Littleton, D.K. Morrison, C. Mungall, P.H. O'Farrell, O.K. Pickeral, C. Shue, L.B. Vosshall, J. Zhang, Q. Zhao, X.H. Zheng, and S. Lewis. 2000. Comparative genomics of the eukaryotes. *Science* **287:** 2204-2215.

Sachs, G.S. 2003. Decision tree for the treatment of bipolar disorder. *J Clin Psychiatry* **64 Suppl 8:** 35-40.

Samudrala, R. and J. Moult. 1997. Handling context-sensitivity in protein structures using graph theory: bona fide prediction. *Proteins* **Suppl 1:** 43-49.

Schwehm, J.M., E.S. Kristyanne, C.C. Biggers, and W.E. Stites. 1998. Stability effects of increasing the hydrophobicity of solvent-exposed side chains in staphylococcal nuclease. *Biochemistry* **37:** 6939-6948.

Scriver, C.R., M. Hurtubise, D. Konecki, M. Phommarinh, L. Prevost, H. Erlandsen, R. Stevens, P.J. Waters, S. Ryan, D. McDonald, and C. Sarkissian. 2003. PAHdb 2003: what a locus-specific knowledgebase can do. *Hum Mutat* **21:** 333-344.

Segre, D., A. Deluna, G.M. Church, and R. Kishony. 2005. Modular epistasis in yeast metabolism. *Nat Genet* **37:** 77-83.

Serrano, L., J.T. Kellis, Jr., P. Cann, A. Matouschek, and A.R. Fersht. 1992a. The folding of an enzyme. II. Substructure of barnase and the contribution of different interactions to protein stability. *J Mol Biol* **224:** 783-804.

Serrano, L., A. Matouschek, and A.R. Fersht. 1992b. The folding of an enzyme. III. Structure of the transition state for unfolding of barnase analysed by a protein engineering procedure. *J Mol Biol* **224:** 805-818.

Serrano, L., A. Matouschek, and A.R. Fersht. 1992c. The folding of an enzyme. VI. The folding pathway of barnase: comparison with theoretical models. *J Mol Biol* **224:** 847-859.

Serrano, L., J. Sancho, M. Hirshberg, and A.R. Fersht. 1992d. Alpha-helix stability in proteins. I. Empirical correlations concerning substitution of side-chains at the N and C-caps and the replacement of alanine by glycine or serine at solvent-exposed surfaces. *J Mol Biol* **227:** 544-559.

Service, R. 2005. Structural biology. Structural genomics, round 2. *Science* **307:** 1554-1558.

Shen, L.X., J.P. Basilion, and V.P. Stanton, Jr. 1999. Single-nucleotide polymorphisms can cause different structural folds of mRNA. *Proc Natl Acad Sci U S A* **96:** 7871-7876.

Sherry, S.T., M.H. Ward, M. Kholodov, J. Baker, L. Phan, E.M. Smigielski, and K. Sirotkin. 2001. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res* **29:** 308-311.

Shinkai, T., O. Ohmori, H. Hori, and J. Nakamura. 2003. Genetic approaches to polydipsia in schizophrenia: a preliminary report of a family study and an association study of an angiotensin-converting enzyme gene polymorphism. *Am J Med Genet B Neuropsychiatr Genet* **119:** 7-12.

Shortle, D., W.E. Stites, and A.K. Meeker. 1990. Contributions of the large hydrophobic amino acids to the stability of staphylococcal nuclease. *Biochemistry* **29:** 8033-8041.

Shrake, A. and J.A. Rupley. 1973. Environment and exposure to solvent of protein atoms. Lysozyme and insulin. *J Mol Biol* **79:** 351-371.

Smith, D., K. McKenna, K. Moore, W. Tormey, J. Finucane, J. Phillips, P. Baylis, and C.J. Thompson. 2002. Baroregulation of vasopressin release in adipsic diabetes insipidus. *J Clin Endocrinol Metab* **87:** 4564-4568.

Smith, D.J. and A.J. Lusis. 2002. The allelic structure of common disease. *Hum Mol Genet* **11:** 2455-2461.

Smith, N.G. and A. Eyre-Walker. 2003. Human disease genes: patterns and predictions. *Gene* **318:** 169-175.

Stapley, B.J. and G. Benoit. 2000. Biobibliometrics: information retrieval and visualization from co-occurrences of gene names in Medline abstracts. *Pac Symp Biocomput***:** 529-540.

Stenson, P.D., E.V. Ball, M. Mort, A.D. Phillips, J.A. Shiel, N.S. Thomas, S. Abeysinghe, M. Krawczak, and D.N. Cooper. 2003. Human Gene Mutation Database (HGMD): 2003 update. *Hum Mutat* **21:** 577-581.

Stites, W.E., A.K. Meeker, and D. Shortle. 1994. Evidence for strained interactions between side-chains and the polypeptide backbone. *J Mol Biol* **235:** 27-32.

Stitziel, N.O., T.A. Binkowski, Y.Y. Tseng, S. Kasif, and J. Liang. 2004. topoSNP: a topographic database of non-synonymous single nucleotide polymorphisms with and without known disease association. *Nucleic Acids Res* **32:** D520-522.

Stitziel, N.O., Y.Y. Tseng, D. Pervouchine, D. Goddeau, S. Kasif, and J. Liang. 2003. Structural location of disease-associated single-nucleotide polymorphisms. *J Mol Biol* **327:** 1021-1030.

Su, A.I., M.P. Cooke, K.A. Ching, Y. Hakak, J.R. Walker, T. Wiltshire, A.P. Orth, R.G. Vega, L.M. Sapinoso, A. Moqrich, A. Patapoutian, G.M. Hampton, P.G.

Schultz, and J.B. Hogenesch. 2002. Large-scale analysis of the human and mouse transcriptomes. *Proc Natl Acad Sci U S A* **99:** 4465-4470.

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

Takai, S. and M. Miyazaki. 2003. Application of a chymase inhibitor, NK3201, for prevention of vascular proliferation. *Cardiovasc Drug Rev* **21:** 185-198.

Tishkoff, S.A. and B.C. Verrelli. 2003. Role of evolutionary history on haplotype block structure in the human genome: implications for disease mapping. *Curr Opin Genet Dev* **13:** 569-575.

Tong, A.H., G. Lesage, G.D. Bader, H. Ding, H. Xu, X. Xin, J. Young, G.F. Berriz, R.L. Brost, M. Chang, Y. Chen, X. Cheng, G. Chua, H. Friesen, D.S. Goldberg, J. Haynes, C. Humphries, G. He, S. Hussein, L. Ke, N. Krogan, Z. Li, J.N. Levinson, H. Lu, P. Menard, C. Munyana, A.B. Parsons, O. Ryan, R. Tonikian, T. Roberts, A.M. Sdicu, J. Shapiro, B. Sheikh, B. Suter, S.L. Wong, L.V. Zhang, H. Zhu, C.G. Burd, S. Munro, C. Sander, J. Rine, J. Greenblatt, M. Peter, A. Bretscher, G. Bell, F.P. Roth, G.W. Brown, B. Andrews, H. Bussey, and C. Boone. 2004. Global mapping of the yeast genetic interaction network. *Science* **303:** 808-813.

Tramontano, A. and V. Morea. 2003. Assessment of homology-based predictions in CASP5. *Proteins* **53 Suppl 6:** 352-368.

Tsirigos, A. and I. Rigoutsos. 2005. A sensitive, support-vector-machine method for the detection of horizontal gene transfers in viral, archaeal and bacterial genomes. *Nucleic Acids Res* **33:** 3699-3707.

Turner, S.T. and E. Boerwinkle. 2003. Genetics of blood pressure, hypertensive complications, and antihypertensive drug responses. *Pharmacogenomics* **4:** 53-65.

Valdar, W.S. and J.M. Thornton. 2001. Conservation helps to identify biologically relevant crystal contacts. *J Mol Biol* **313:** 399-416.

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

Venter, J.C. M.D. Adams E.W. Myers P.W. Li R.J. Mural G.G. Sutton H.O. Smith M. Yandell C.A. Evans R.A. Holt J.D. Gocayne P. Amanatides R.M. Ballew D.H. Huson J.R. Wortman Q. Zhang C.D. Kodira X.H. Zheng L. Chen M. Skupski G. Subramanian P.D. Thomas J. Zhang G.L. Gabor Miklos C. Nelson S. Broder A.G. Clark J. Nadeau V.A. McKusick N. Zinder A.J. Levine R.J. Roberts M. Simon C. Slayman M. Hunkapiller R. Bolanos A. Delcher I. Dew D. Fasulo M. Flanigan L. Florea A. Halpern S. Hannenhalli S. Kravitz S. Levy C. Mobarry K. Reinert K. Remington J. Abu-Threideh E. Beasley K. Biddick V. Bonazzi R. Brandon M. Cargill I. Chandramouliswaran R. Charlab K. Chaturvedi Z. Deng V. Di Francesco P. Dunn K. Eilbeck C. Evangelista A.E. Gabrielian W. Gan W. Ge F. Gong Z. Gu P. Guan T.J. Heiman M.E. Higgins R.R. Ji Z. Ke K.A. Ketchum Z. Lai Y. Lei Z. Li J. Li Y. Liang X. Lin F. Lu G.V. Merkulov N. Milshina H.M. Moore A.K. Naik V.A. Narayan B. Neelam D. Nusskern D.B. Rusch S. Salzberg W. Shao B. Shue J. Sun Z. Wang A. Wang X. Wang J. Wang M. Wei R. Wides C. Xiao C. Yan A. Yao J. Ye M. Zhan W. Zhang H. Zhang Q. Zhao L. Zheng F. Zhong W. Zhong S. Zhu S. Zhao D. Gilbert S. Baumhueter G. Spier C. Carter A. Cravchik T. Woodage F. Ali H. An A. Awe D. Baldwin H. Baden M. Barnstead I. Barrow K. Beeson D. Busam A. Carver A. Center M.L. Cheng L. Curry S. Danaher L. Davenport R. Desilets S. Dietz K. Dodson L. Doup S. Ferriera N. Garg A. Gluecksmann B. Hart J. Haynes C. Haynes C. Heiner S. Hladun D. Hostin J. Houck T. Howland C. Ibegwam J. Johnson F. Kalush L. Kline S. Koduru A. Love F. Mann D. May S. McCawley T. McIntosh I. McMullen M. Moy L. Moy B. Murphy K. Nelson C. Pfannkoch E. Pratts V. Puri H. Qureshi M. Reardon R. Rodriguez Y.H. Rogers D. Romblad B. Ruhfel R. Scott C. Sitter M. Smallwood E. Stewart R. Strong E. Suh R. Thomas N.N. Tint S. Tse C. Vech G. Wang J. Wetter S. Williams M. Williams S. Windsor E. Winn-Deen K. Wolfe J. Zaveri K. Zaveri J.F. Abril R. Guigo M.J. Campbell K.V. Sjolander B. Karlak A. Kejariwal H. Mi B. Lazareva T. Hatton A. Narechania K. Diemer A. Muruganujan N. Guo S. Sato V. Bafna S. Istrail R. Lippert R. Schwartz B. Walenz S. Yooseph D. Allen A. Basu J. Baxendale L. Blick M. Caminha J. Carnes-Stine P. Caulk Y.H. Chiang M. Coyne C. Dahlke A. Mays M. Dombroski M. Donnelly D. Ely S. Esparham C. Fosler H. Gire S. Glanowski K. Glasser A. Glodek M. Gorokhov K. Graham B. Gropman M. Harris J. Heil S. Henderson J. Hoover D. Jennings C. Jordan J. Jordan J. Kasha L. Kagan C. Kraft A. Levitsky M. Lewis X. Liu J. Lopez D. Ma W. Majoros J. McDaniel S. Murphy M. Newman T. Nguyen N. Nguyen M. Nodell S. Pan J. Peck M. Peterson W. Rowe R. Sanders J. Scott M. Simpson T. Smith A. Sprague T. Stockwell R. Turner E. Venter M. Wang M. Wen D. Wu M. Wu A. Xia A. Zandieh and X. Zhu. 2001. The sequence of the human genome. *Science* **291:** 1304-1351.

Waltz, M.R., T.N. Pullman, K. Takeda, P. Sobieszczyk, and S. Refetoff. 1990. Molecular basis for the properties of the thyroxine-binding globulin-slow variant in American blacks. *J Endocrinol Invest* **13:** 343-349.

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

Wang, Z. and J. Moult. 2003. Three-dimensional structural location and molecular functional effects of missense SNPs in the T cell receptor Vbeta domain. *Proteins* **53:** 748-757.

Wheeler, D.L., T. Barrett, D.A. Benson, S.H. Bryant, K. Canese, D.M. Church, M. DiCuccio, R. Edgar, S. Federhen, W. Helmberg, D.L. Kenton, O. Khovayko, D.J. Lipman, T.L. Madden, D.R. Maglott, J. Ostell, J.U. Pontius, K.D. Pruitt, G.D. Schuler, L.M. Schriml, E. Sequeira, S.T. Sherry, K. Sirotkin, G. Starchenko, T.O. Suzek, R. Tatusov, T.A. Tatusova, L. Wagner, and E. Yaschenko. 2005. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* **33:** D39-45.

Wheeler, D.L., D.M. Church, R. Edgar, S. Federhen, W. Helmberg, T.L. Madden, J.U. Pontius, G.D. Schuler, L.M. Schriml, E. Sequeira, T.O. Suzek, T.A. Tatusova, and L. Wagner. 2004. Database resources of the National Center for Biotechnology Information: update. *Nucleic Acids Res* **32:** D35-40.

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

Yue, P., Moult, J. 2005. Identification and Analysis of Deleterious Human SNPs. *Submitted*.

Zhao, C.Y., H.X. Zhang, X.Y. Zhang, M.C. Liu, Z.D. Hu, and B.T. Fan. 2005. Application of support vector machine (SVM) for prediction toxic activity of different data sets. *Toxicology*.

Glossary

ANOVA:
Analysis of variance (ANOVA) is a test of the statistical significance of the differences among the mean scores of two or more groups.

Backbone Strain:
An unfavorable contribution to energy of the folded configuration arising from close atomic contact. Backbone strain can be caused in three different ways: replacement of a glycine residue with φ/ψ angles in a non-allowed region for other residues, replacement of cis-proline with another residue, and replacement of another residue by proline where the φ value is inappropriate. Mutagenesis data show that backbone strain may result in up to 2 Kcal/mol loss of free energy of stabilization.

B factor:
See Crystallographic B-factor.

BIND:
The Biomolecular Interaction Network Database (BIND) is a database of biomolecular interactions.

BLAST:
BLAST (Basic Local Alignment Search Tool) is a computational method for rapid searching of nucleotide and protein databases for sequences similar to a query sequence. An amino-acid substitution matrix is used by BLAST to calculate the sequence similarity score between sequences. The default matrix of BLAST is BLOSUM62.

BLOSUM
BLOSUM (Blocks Substitution Matrix) is a type of amino-acid substitution matrix. It is derived from sequence alignments within conserved protein families. The frequency of each amino acid substitution is calculated from the alignments. Different levels of the BLOSUM matrix can be created from different levels of sequence similarity. For example, the BLOSUM62 matrix is calculated from protein blocks where no two sequences are more than 62% identity.

Breakage of a disulfide bond:
In a protein, a disulfide bond is the bond between a pair of Cysteine residues. Breakage of a disulfide bond by mutating one Cysteine to a different residue usually has a large effect on protein stability.

Buried charge:
An unpaired charged residue introduced into the hydrophobic core of a protein by mutation. Buried charged residues are known to destabilize proteins by 3-5 Kcal/mol.

Buried Polar:
An unpaired polar residue introduced into the hydrophobic core of a protein by mutation. A buried polar residue is known to be destabilizing.

CASP:
CASP (Critical Assessment of Techniques for Protein Structure Prediction) is a community-wide experiment to make blind predictions on the structures of a set of proteins whose solved structures are temporally hidden from participants. The goal is to establish the current state of the art in protein structure prediction.

Cα
The backbone aliphatic carbon atom of an amino acid is called Cα, and is bonded to an amino group, a carboxyl group, the side chain and one hydrogen atom.

Cavity:
A **cavity** is an interior empty space in a protein structure. A cavity can be created by mutating a residue to one with a smaller side chain in the core of a protein, and is known to be destabilizing to a protein structure.

CD/CV
The common disease-common variant model. It assumes that common diseases are affected by common disease-susceptibility alleles at a small number of loci that exist at a high frequency across populations.

CLUSTALW:
CLUSTALW is a multiple sequence alignment program for DNA or proteins.

Comparative Modeling:
Comparative Modeling, also termed homology modeling, is a method which is used to model a three-dimensional structure from the structures of homologous proteins.

Concept Profile:
A concept profile refers to an ordered list of terms that are most closely associated with the concept of interest in the literature.

Crystallographic B-factor: (also referred as B factor)
Atomic B factors are obtained from the crystallographic refinement of protein structures, and are a measure of the diffuseness of the electron density distribution around atoms. A high B factor indicates the relatively high mobility of the corresponding atom. It has been suggested that regions with high B factors tend to be more tolerant of mutations.

dbSNP:
dbSNP is a central, public, repository for SNP data.

DIP:
The DIP^{TM} (database of interacting proteins) catalogs experimentally determined interactions between proteins.

EcoCyc:
EcoCyc is a database for the *Escherichia coli* K-12 MG1655 bacterium. The database contains a range of curated biological information, such as transcriptional regulation and metabolic pathways.

Electrostatic interaction:
Protein structures are organized such that almost all polar and charged groups are in locally favorable electrostatic environments. We divide electrostatic interactions into three types: hydrogen-bond between polar-polar groups (PP), hydrogen-bond between polar-charge groups (PC), and saltbridge between charge-charge groups (CC). *In vitro* mutagenesis data show that removing a hydrogen-bond or salt-bridge will destabilize protein structure.

Electrostatic repulsion:
Electrostatic repulsion is the repulsion between two same charges. Electrostatic repulsion is known to destabilize a protein structure.

Epistasis:
Epistasis is non-linear interaction between genes affecting a single phenotype.

Entropy:
Entropy is a quantity used to measure the degree of disorder in a system. The higher the entropy, the greater the disorder.

E-score:
Expect value, also termed E-value. The E-score is a parameter that is used by BLAST and PSIBLAST to describe the chance by which a sequence similarity hit can be seen when searching a sequence database. The lower the E-score is, the lower the chance.

FN:
FN is the number of false negatives in a test.

FOLD-X:
FOLD-X is a program for calculating the effect of a single residue mutations on the stability of a protein.

FP:
FP is the number of false positives in a test.

GAD:

The Genetic Association Database (GAD) is an archive of human genetic association studies of complex diseases and disorders.

GO:
The Gene Ontology (GO) is a database that provides a controlled vocabulary to describe gene and gene product attributes in any organism.

HGMD:
The Human Gene Mutation Database (HGMD) is a database of published gene lesions responsible for human inherited disease (mostly monogenic disease).

Homolog:
Homologs are genes that are descendent from the same ancestor. Paralogs and orthologs are two forms of homologs.

HomoloGene:
HomoloGene is a NCBI database that collects homologs among the annotated genes of several completely sequenced eukaryotic genomes.

HSSP:
The HSSP is a database of homology-derived secondary structure of proteins.

Hydrophobic burial:
The hydrophobic effect is considered to be the major driving force for the folding of globular proteins. It causes nonpolar side-chains to cluster in proteins. The non-polar area buried in a folded protein is used to quantify hydrophobic burial.

Jmol:
Jmol is a free software package that is used to view three-dimensional structures.

KEGG
KEGG (Kyoto Encyclopedia of Genes and Genomes) is a suite of databases and associated software which facilitate integration of the current knowledge on biological information.

LD
Linkage Disequilibrium (LD) is the non-random association between genetic markers in a population.

Machine learning:
Machine learning refers to a system that is capable of autonomous acquisition and integration of knowledge. It usually requires training an algorithm on a given data set and testing it on other data set.

Mis-sense SNPs:

Mis-sense SNPs, also termed non-synonymous SNPs, are SNPs that are located in coding regions and result in amino acid variation in the protein products of genes.

MySQL:
MySQL is a database management system.

Natural language processing (NLP):
Natural Language Processing (NLP) technology is software for analyzing, understanding and generating natural human language.

Non-synonymous SNPs:
See mis-sense SNPs.

NR:
NR is the NCBI non-redundant protein sequence database.

OMIM
Online (Mendelian Inheritance in Man) is a database of human genes and genetic disorders.

Ortholog:
Orthologs are genes in different species that evolved from a common ancestral gene by speciation.

Overpacking
Introducing a residue with a large side chain into the core of a protein may cause steric clashes between this residue and the surrounding residues. This phenomenon is called overpacking and destabilizing protein structure.

Paralog:
Paralogs are genes related by duplication within a genome.

PDB:
PDB (Protein Data Bank) is the database of the 3-D structures of proteins and nucleic acids.

PHD:
PHD is a program for predicting protein secondary structure and per residue solvent accessibility from multiple sequence alignments.

Phenylketonuria (PKU):
Phenylketonuria (PKU) is a genetic disorder that is characterized by an impaired ability to process phenylalanine into other compounds.

PHP:
PHP is a scripting language that has been widely used for web development.

PHP-NUKE:
PHP-NUKE is an open source template for web development. It is written in PHP.

PQS:
Protein Quaternary Structure (PQS) is a database of probable protein quaternary structures based on structures in the PDB database.

Principal component analysis
Principal component analysis (PCA) is a mathematical procedure that transforms a number of (possibly) correlated variables into a (smaller) number of uncorrelated variables called principal components.

PROCHECK:
PROCHECK is a program for assessing the stereochemical quality of a given protein structure.

Protein Stability:
The stability of a protein is the difference in Gibbs free energy (ΔG) between the folded and the unfolded states of a protein.

ProTherm:
ProTherm(Protein Thermodynamic Database) is a database that collects numerical data of thermodynamic parameters (such as Gibbs free energy change and enthalpy change on folding) for wild type and mutant proteins.

PSIBLAST
PSIBLAST (Altschul et al. 1997) is a program that iteratively searches protein databases for sequences similar to the query sequence. PSIBLAST and BLAST are similar except that the former uses position-specific scoring matrices (PSSMs) generated in the searching process while the later uses pre-defined substitution matrices such as BLOSUM62.

PSIBLAST can be used to repeatedly search target databases. It uses a multiple alignment of high scoring sequences found in each search round to generate a new PSSM for use in the next round of searching. PSIBLAST will iterate until no new sequences are found, or the user may specify a maximum number of iterations. A maximum of three iterations is used in the profile model.

PSSM
PSSM is a position-specific scoring matrix. It is generated by PSIBLAST during the process of searching for the sequences related to a query sequence.

RefSeq:
RefSeq is the NCBI database of reference sequences. It contains a curated and non-redundant set of nucleotide and protein sequences.

RefSNP:
RefSNP is a non-redundant set of variations in dbSNP.

Relative Surface accessibility:
See surface accessibility.

Root mean square (RMS) error:
In this work, the root mean square (RMS) error is used to measure the distance between two 3-dimensional structures. The RMS error is defined as the root mean square distance between sets of related atoms.

SCOP:
Structural Classification of Proteins (SCOP) is a database of protein structure classification.

ScoreCons:
A program for scoring residue conservation in a multiple sequence alignment.

SCWRL:
SCWRL is a program for adding sidechains to a protein backbone based on a backbone-dependent rotamer library. The library provides lists of χ1-χ2-χ3-χ4 values and their related probabilities for residues with given φ-ψ values. The library is generated from a selected list of solved protein structures.

Sensitivity:
Sensitivity = TP/TP+FN, where TP is the number of true positives and FN is the number of false negatives.

Sequence Profile:
A Sequence Profile in this dissertation is defined as a multiple sequence alignment between a human sequence and its homologs.

SNPSubSNPLink:
SNPSubSNPLink is a mapping table of RefSNP IDs and the corresponding Submitted SNP IDs.

Specificity:
Specificity = TN/TN+FP, where TN is the number of true negatives and FP is the number of false positives.

(Relative) Surface accessibility (or solvent accessibility):
Solvent surface area describes the area of a protein that is accessible to solvent. In order to calculate the solvent surface area of a protein or residue, a probe sphere representing the solvent molecule is rolled over the protein surface. The contact surface between the protein molecule (solute) and the solvent molecule is defined as

the solvent surface area. The surface accessibility of a residue is represented by the ratio between the solvent surface area of the residue in a folded protein and that in an unfolded protein. A residue is classified as on the protein surface if its surface accessibility is more than 20%.

SVM
SVM is a computational method of data classification.

Swiss-Prot:
Swiss-Prot is a curated protein sequence database.

Temperature factor:
See Crystallographic B-factor.

Z-score:
Z-score is a statistical measure that quantifies the difference (measured in standard deviations) between a sample and the mean of a data set.