Technische Universität München

Lehrstuhl für Bioinformatik

Prediction of functional effects for sequence variants

Maximilian Natale Friedrich Hecht

Vollständiger Abdruck der von der Fakultät für Informatik der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften

genehmigten Dissertation.

Vorsitzender:	Univ.-Prof. Dr. Daniel Cremers
Prüfer der Dissertation:	
	1. Univ.-Prof. Dr. Burkhard Rost
	2. Univ.-Prof. Dr. Dimitri Frischmann

Die Dissertation wurde am 24.06.2015 bei der Technischen Universität München eingereicht und durch die Fakultät für Informatik am 13.10.2015 angenommen.

Contents

Abstract

Technological advancements have enabled us to find and record genetic differences between humans. Yet, the vast amount of data gathered through next generation sequencing has long since overwhelmed our ability to study every discovered difference experimentally. Thus, computational approaches have been developed to cope with the deluge of data and guide experimental efforts towards prioritizing the most promising candidates. This thesis focuses on the computational advancements made towards predicting the effects of genetic variation. It discusses the current state of research and presents a newly developed method for the prediction of functional effects of sequence variants. Our new method SNAP2 predicts variants with over 83% accuracy and is also able to predict effects for variants of orphan proteins. Both constitute significant and important improvements over other methods. Furthermore, it not only predicts single variants but offers a comprehensive view of all possible substitution effects in a protein. This opens up a novel view on the landscape of protein mutability. Possible applications are presented that show how this novel view on functional effect predictions can translate into novel hypotheses and thus aid the identification of targets for drug development.

1 Introduction

Understanding human genetic variation is one of the major scientific challenges of the 21st century. Nearly 15 years after we first sequenced the human genome, we still understand little of the mechanisms that link differences on the genetic level to the differences observed on the phenotypic level. Genetic variations influence phenotypes: from the most obvious, such as the visible differences between individuals with different ethnic background, to the least obvious such as the differential response to drug treatment. Therefore, the genotype-phenotype link is crucial for understanding development and progression of diseases: it may be the key towards developing personalized treatment.

1.1 Forms of genetic variation

Differing inheritable traits can either favor or hinder survival and reproduction. This process is called natural selection. The interplay between genetic variation and selection pressure is what constantly adapts organisms to best fit their specific niche in the environment. In nature genetic variation happens randomly through changes on the molecular level which are caused by various factors, such as mutations and random mating between organisms (Krishnamurthy, 2003). There are different forms of genetic variation, which can be categorized according to their size and type: (i) numerical variation, (ii) large-scale structural variation and (iii) small-scale sequence variation.

Numerical Variation

Numerical variation can be defined as changes in the number of chromosomes, referred to as polyploidy (numerical change of the whole set of chromosomes) or aneuploidy (number of individual chromosomes is altered). A prominent example of polyploidy is wheat, which through years of hybridization and human modification has different species that range from diploid (two sets

of chromosomes) to hexaploid (six sets of chromosomes; today's common bread wheat) (Martinez-Perez et al., 2003). In humans polyploidy plays not only a role in the normal development of certain cell types but also in the development of cancer (Davoli and de Lange, 2011). For humans, aneuploidy of most chromosomes results in miscarriage but there are exceptions that result in live births (Driscoll and Gross, 2009). These children often suffer from genetic disorders, the most common example being trisomy 21 (known as Down syndrome), where a third (possibly partial) copy of chromosome 21 is present (Patterson, 2009). Aneuploidy can also be observed in cancer cells (Sen, 2000).

Large-scale structural variation

The class of large-scale structural variation comprises variations of long stretches/regions of DNA, typically ranging in size from kilo- to megabases. These structural variations include rearrangements such as translocations (exchange of region between chromosomes) or inversions (region is reversed in the chromosome) but also so-called Copy-Number Variations (CNV; region is duplicated or deleted). While balanced rearrangements (reciprocal translocations or inversions) do not result in gain or loss of genetic material they may still affect gene expression through gene fusion or by dissociating genes from their long-range regulatory elements (Harewood et al., 2010). CNVs, on the other hand, represent a significant alteration of the genetic material resulting from duplication or deletion of long stretches of chromosomes. These large-scale variations were found to be widespread and common among humans (Iafrate et al., 2004; Sebat et al., 2004) accounting for roughly 13% of the human genome (Stankiewicz and Lupski, 2010).

Yet, with respect to CNVs only 0.4% of the genome significantly differed in a study comparing eight individual genomes to the reference assembly (Kidd et al., 2008). This suggests that most of the CNVs are common and shared by members of the same population. While CNVs have been associated with

diseases, for instance autism and schizophrenia (Cook and Scherer, 2008), it appears that gene gains are more common than gene losses and that gene amplification is favored through positive selection (Zhang et al., 2009). For example, the salivary amylase gene (AMY1) copy number was found to be varying significantly in different populations. This was considered an adaption of agricultural societies towards high-starch diet, as higher copy numbers correlated with higher salivary starch-digesting enzyme levels (Perry et al., 2007).

Small-scale sequence variation

The last class, the small-scale sequence variation, is the most prevalent form of genetic variation among humans (ENCODE Project Consortium, 2012). It comprises short or single nucleotide insertions/deletions (typically abbreviated to 'indels') and single nucleotide substitutions (so-called point mutations). Indels (that are not a multiple of 3 bases) in genetic regions that encode proteins or functional RNA change the reading frame. This is problematic because it affects how the codon triplets are read during transcription into mRNA (or other RNA). These so-called frameshift mutations have been shown to be involved in a number of diseases: For instance, the Crohn's disease has been associated with an insertion in the NOD2 gene. The insertion of cytosine at position 3020 of the NOD2 gene was shown to produce a truncated protein. The resulting protein no longer responded to bacterial lipopolysaccharides, which might lead to increased susceptibility to Crohn's disease (Ogura et al., 2001).

Single nucleotide variation

The most frequent human genetic variations are single nucleotide substitutions (called 'single nucleotide polymorphisms' – SNPs, or 'single nucleotide variants' – SNVs). The international SNP Map Working Group estimated that any two haploid genomes differed on average by around 1 nucleotide

every 1300 basepairs (Sachidanandam et al., 2001), a number which likely varies between ethnic groups. Estimates suggest that the human genome contains over 11 million SNPs of which roughly 7 million are common (occurring at a minor allele frequency [MAF] greater than 5%) in the human population while the remaining 4 million are uncommon ($1\% < \text{MAF} \leq 5\%$) (Kruglyak and Nickerson, 2001). These SNPs are estimated to constitute 90% of the genetic variation in humans while the remaining 10% consist of a vast array of variants that are each rare within the population (International HapMap Consortium, 2003). Researchers suspect to find a tremendous amount of these rare variants. It is likely that almost every 'life-compatible' variant can be observed in at least one of the roughly 7 billion people on earth (Frazer et al., 2009). The vast majority of SNPs is located in non-protein-coding regions of the genome, since only around 1.5% of the genome encode proteins (Lander et al., 2001). Although these SNPs do not directly alter gene products they may for instance affect gene regulation by changing transcription factor binding or gene splicing. A more direct phenotypic impact can be expected for coding SNPs (cSNPs). Due to the degeneracy of the genetic code, SNPs in coding regions can either be synonymous (sSNPs) or non-synonymous (nsSNPs). The former change the codon triplets in such a way that the encoded amino acid is the same as before and thus do not alter the gene product. The latter, however, change the codon triplet in two ways. Either it becomes a stop codon (so-called nonsense mutations) or the codon encodes a different amino acid (so-called missense mutations). While nonsense mutations cause truncated and thus mostly nonfunctional proteins (depending on where the new stop codon is located), missense mutations can have a variety of effects on the resulting protein: They may affect folding and structure of the protein and thus affect its function. They may also change residues that are important for binding and thus affect interactions with substrates or other proteins (*e.g.* complex formation).

These protein-altering point mutations are of particular importance for

medical research because they can directly affect phenotypes such as causing diseases, increasing disease susceptibility or altering drug response (Thusberg and Vihinen, 2009). Although these appear to be a small fraction of variants, there is an estimated average of 6 coding SNPs per gene in the human population (Collins et al., 1998). Recent studies suggest that any two unrelated individuals on average differ by one cSNP per gene, half of which are non-synonymous (1000 Genomes Project Consortium et al., 2010). Thus, approximately every other protein differs between any two individuals from the same population and many of these differences are likely instrumental in defining human diversity.

1.2 Investigating human genetic variation

While genetic variation works on genotypes, natural selection works on phenotypes. Selection is likely to favor (or to not penalize) genetic changes that do not lead to disadvantageous phenotypes. However, not all genetic changes lead to phenotypic changes. In fact, the majority of genetic variation is hypothesized to be neutral (Kimura, 1968) with respect to any phenotypes, *i.e.* they are assumed not to contribute to any phenotype. Yet, the exact ratio of neutral, 'near-neutral' (Ohta, 2002) and non-neutral variation is unknown. Nevertheless, near-neutral variants may contribute (although with little impact each) to complex traits and even non-synonymous neutral variants might be important for human individuality (Bromberg et al., 2013).

As mentioned before, the vast majority of human genetic variation is due to common variants. In other words, the majority of variants in any given individual are variants that are common within the whole population. Moreover, when two individual genomes are compared, the vast majority of differences can be found at positions that are commonly known to be variable within the population (Frazer et al., 2009). Therefore a great effort has been made towards cataloging and studying common variation. One observation towards this end has been particularly important: SNPs in proximity of

one another on the chromosome are likely to be co-inherited, thus leading to a strong correlation between SNPs in the same genomic interval. This complex correlation structure is called linkage disequilibrium (LD) and varies between populations (Slatkin, 2008). The International HapMap Project determined that 80% of common SNPs (MAF $>$ 5%) could be categorized into roughly 550,000 LD groups for individuals of European or Asian ancestry and roughly 1,100,000 LD groups for African ancestry (International HapMap Consortium, 2003). This means that information for over 80% of common SNPs can be gained only by genotyping individual DNA with 'tag' SNPs from each LD group (Barrett and Cardon, 2006; Eberle et al., 2007; Pe'er et al., 2006; Frazer et al., 2009). This remarkable finding significantly reduced the cost and time required to genotype thousands of individuals and thus enabled large-scale genotype-phenotype association studies.

In genome-wide association studies (GWAS) large amounts of participants are genotyped and assigned to either a case or a control group depending on the phenotype under investigation. However, the power of GWAS is limited due to several reasons (Frazer et al., 2009): (i) Study cases need to be representative and sufficient in number, often requiring over 10,000 samples for detection (Kiezun et al., 2012). Moreover, participants are typically drawn from clinical sources, which often do not contain silent, mild or lethal cases because they do not come to clinical attention. (ii) GWAS are limited to common variants, as rare variants are generally not tagged. (iii) 20% of the common variants are not or only partially tagged. Nevertheless, these studies have significantly furthered our understanding of variants and diseases. Most common variants have been tested for associations with common traits and diseases thereby linking more than 1,100 loci to complex diseases (Lander, 2011). Yet, due to their limitations, GWAS cannot find significant associations for rare variants and moreover miss known associations for many common SNPs (Kiezun et al., 2012).

Another approach for detection and investigation of variants and their ef-

fects are so-called Trio Studies. These typically focus on high-coverage whole genome sequencing of mother-father-child trios with the aim of discovering very rare *de novo* variants (*i.e* mutations that cannot be found in either parent). Using this study design, the 1000 Genomes Consortium estimated the *de novo* germ-line mutation rate to be 10^{-8} substitutions per base per generation (1000 Genomes Project Consortium et al., 2010) - an approximate average of 30 new mutations (and thus a theoretical average of roughly $30 * 0.015 = 0.45$ protein coding variants) in each newborn that are not observable in either parent. A different study, however, found 1.2-1.7 coding variants to be novel with respect to both parents (Rauch et al., 2012).

A complementary approach is provided through exome sequencing studies. These studies are aimed at comprehensively assessing both common and rare coding variants by targeted sequencing of the protein coding regions of the genome rather than sequencing the whole genome or relying on LD patterns as in GWAS. Exome-sequencing profits from the accuracy and coverage of Next Generation Sequencing combined with efficient DNA capturing while focussing on a critical region of the genome (Teer and Mullikin, 2010; Kiezun et al., 2012). This allows for significantly larger sample sizes than currently feasible for whole genome sequencing but is limited to the detection of coding SNPs.

This strategy has also been employed in the 1000 Genomes Project (1000 Genomes Project Consortium et al., 2010, 2012) in order to record detailed information on the variants present in 1,092 healthy individuals from 14 populations. Among their astonishing results were many important findings: Each individual carries 80-100 variants causing pre-mature stop codons, 40-50 splice-site disrupting variants, and 220-250 frameshift mutations affecting roughly 250-300 genes. In other words, everyone carries up to 400 variants that are likely to completely disrupt protein function (*i.e.* putative loss-of-function variants). Moreover, the 1000 Genomes Project revealed that every individual also carries an average of 50-100 variants that had previously been

associated with inherited disorders. This clearly shows that the genotype-phenotype relationship is very complex in most cases and that we need to learn more about the underlying molecular mechanisms that link genetic variation to complex traits and diseases. Large-scale exome sequencing studies give valuable insights on the statistical involvement of coding variants in a certain phenotype (*i.e.* typically a certain disease). However, they do not answer how these particular variants affect protein functions, their pathways or the interactions that finally lead to the observable phenotype.

1.3 Resources for variants and their molecular effects

Improved sequencing and variant calling methods have led to a flood of genetic variation being discovered over the last decade. The need of storing these data and making them publicly available gave rise to a variety of databases that collect, store, and present this variation depending on the database's focus. The largest database is dbSNP (Sherry et al., 2001). It collects not only SNPs but also small-scale indels, retroposable element insertions and short tandem repeats along with the corresponding sequence context, its frequency, and additional submission data. It does, however, not provide information on the molecular effect of variants but links to additional sources if phenotypic information is available. The current build (142, Oct. 2014) comprises over 112 million human variants (so-called RefSNP Clusters) of which almost 54 million are located in coding regions.

The Online Mendelian Inheritance in Men (OMIM) database focusses on diseases with a known genetic component (Hamosh et al., 2005). It provides literature-derived information on mendelian phenotypes and also lists associated SNPs for over 3,400 entries (Feb. 2015). A similar approach is provided by the Human Gene Mutation Database (HGMD; Stenson et al., 2003). This database collects data on germ-line mutations in genes associated with human diseases and currently covers over 64,000 publicly available missense and nonsense variants (Feb. 2015). Another disease-association based repos-

itory is provided through GWAS Central (Beck et al., 2014). This database contains summary level information from genome-wide association studies, providing almost 68 million association p-values for almost 3 million unique dbSNP markers (release 11, Sep. 2013). In addition to these, the Universal Protein Resource (UniProt; UniProt Consortium, 2015) provides an index (called HUMSAVAR) for all variant entries with disease association, covering roughly 70,000 coding variants (release 2015_2 of Feb. 2015). All these resources (OMIM, HGMD, GWAS, HUMSAVAR) offer information on variants with probable disease association based on literature reports but generally do not provide information on molecular variant effects. It is however likely that many of the disease-associated coding variants have functional effects on the molecular level. This will be discussed later in section 2.1.

Studying variant effects experimentally on the molecular level is costly and requires significant amounts of time. Effects are mostly investigated for non-synonymous coding variants because these variants change the protein product and are therefore most likely to have phenotypic effects. The most common way of testing molecular effects of coding variants is site-directed mutagenesis in combination with a certain assay: the effect of a certain substitution is measured with respect to the specified assay. In most cases only a few possible variants are tested experimentally. Exceptions are the comprehensive mutagenesis study of the *E. coli* LacI repressor (Markiewicz et al., 1994), in which over 4,000 variants of the lac repressor were tested (Fig. 1) or the complete mutagenesis of the HIV-I protease (Loeb et al., 1989).

Studies of structural (impact on the native 3D protein structure) and functional effects were collected into a literature-derived database termed the Protein Mutant Database (PMD; Kawabata et al., 1999), whose latest build (March 2007) contains over 200,000 variant entries for roughly 45,000 proteins. Yet, mutagenesis studies are often not aimed at investigating the effect of mutations but rather aimed at studying the native protein function

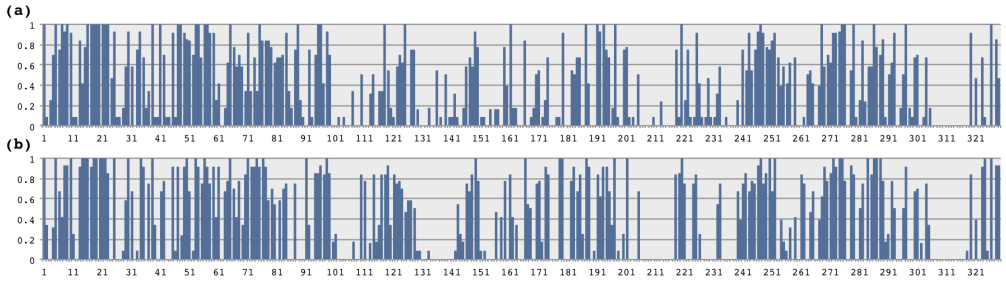

Figure 1: Mutagenesis of the *E. coli* lacI repressor. At each position between residue 2 and 329, 12-13 amino acid substitutions are displayed as a bar. The height of a bar depicts the relative percentage of substitutions that alter the repressor function as determined **(a)** experimentally (Markiewicz et al., 1994) or **(b)** by computational prediction using SNAP2. This figure was adapted from Hecht et al. (2013).

and identifying the residues that are involved. For instance, alanine-scans are often used to probe the protein for functional hot spots by mutating every native amino acid into alanine and testing its effect on protein binding (Bogan and Thorn, 1998). Binding hot spots can be revealed by measuring the difference in binding free energy upon mutation to alanine (Clackson and Wells, 1995). Thus, such experiments also generate information on the molecular effects of to-alanine substitutions in these proteins which is collected in the Alanine Scanning Energetics database (ASEdb; Thorn and Bogan, 2001).

The presented databases constitute only a fraction of the available data repositories and have been presented because of their particular importance to this thesis. Yet, there are many other databases that provide invaluable sources of information for bioinformatics applications.

1.4 Prediction of variant effects

While next-generation sequencing technology becomes increasingly cheaper and faster, experimental verification of variants remains a bottleneck. Hence, fast and accurate computational prediction of variant effects becomes more

and more important in order to guide and prioritize experimental verification of the ever-increasing deluge of sequencing data. In the following, the term variant is used as a synonym for 'coding SNP causing a single amino acid substitution', while non-coding SNPs will be mostly neglected for simplicity. Many methods for the prediction of variant effects have been developed in the past 15 years.

Some methods (*e.g.* DISCERN: Sankararaman et al., 2010, INTERPID: Sankararaman and Sjölander, 2008) are aimed at finding active sites like the ones that significantly alter binding energy when mutated to alanine as annotated in ASEdb. By looking for evolutionary conserved residues these methods also capture a significant fraction of residues that, if mutated, cause disorders or other distinct phenotypes. While not being specifically optimized towards variant prediction, these methods can be used to predict that most variants of these (often highly conserved) active site residues will have a strong impact on the protein function.

Methods specifically trained on variants cover the prediction of a variety of effect aspects, including the explicit prediction of changes in protein-protein binding affinity upon mutation (*e.g.* BeAtMuSiC: Dehouck et al., 2013). Some predict the pathogenicity of coding SNPs (*e.g.* CADD: Kircher et al., 2014, SNPs&GO: Calabrese et al., 2009, Mutation Taster: Schwarz et al., 2010, Mutation Assessor: Reva et al., 2011, PolyPhen-2: Adzhubei et al., 2010) and non-coding SNPs (CADD, Mutation Taster), in the sense that they output a score reflecting the likelihood of a specific SNP being deleterious. Others focus on predicting effects of variants on protein structure (Schaefer and Rost, 2012) and stability (*e.g.* i-Mutant-3: Capriotti et al., 2008, PoPMuSiC: Dehouck et al., 2009).

Others yet put their focus on predicting whether or not a variant changes the native protein function (*e.g.* SNAP: Bromberg and Rost, 2007, SIFT: Kumar et al., 2009, PolyPhen-2: Adzhubei et al., 2010). Obviously, there can be significant overlap between these methods' predictions since the effect

aspects are related. For instance, a variant that causes a structural change in the protein may reduce its function and thus cause a certain disease. However, a pairwise comparison of methods suggests that predictions vary considerably between methods, even within the same category, which will be discussed in section 3.

Prediction features

The focus of each method is essentially determined by the features used for prediction and the data used for training. The single most important feature for variant effect prediction is typically evolutionary conservation - the extent to which the native amino acid is conserved between homologous sequences. If the native amino acid is identical in most (especially also distantly) related species, it is likely that this is due to functional importance as the amino acid was retained through purifying selection (*i.e.* variants causing a disadvantageous phenotype are selected against). This can sometimes be misleading if the alignment does not contain enough distantly related sequences, as the apparent conservation may also result from a lack of time (*i.e.* closely related species did not have enough time to diverge). This feature therefore requires a large and sufficiently diverse alignment of homologous sequences, which is sometimes not available because the required sequences are unknown. Another common feature are biophysical properties of native and variant amino acid. Information such as size, hydrophobicity and charge of the variant is compared with the native amino acid to estimate substitution compatibility. For instance, substituting a small, hydrophobic amino acid in the protein core by a bulky, hydrophilic one is likely to cause a structural change of the protein. Thus, the extent to which biophysical properties differ can be used as an indicator of variant effect. Moreover, many other features of protein and amino-acid features, both experimental and predicted, have been used for variant effect predictions. These include structural information such as experimental structures or predicted secondary structures, residue annota-

tion such as known or predicted functional sites, patterns of protein domains and substitution probabilities estimated from known sequences. Features for variant effect prediction will be further discussed in section 2.2.

Classes of predictors

Predictors can usually be classified into two categories: constraint-based methods and trained classifiers. Constraint-based methods use a custom definition for their predicted effect aspect, which is based on one or more biological properties. For instance, a simple definition for "disease-variant" could be 'over n% residue conservation' and 'm% different biophysical amino-acid properties'. The exact definition parameters (here, 'n' and 'm') are typically optimized using available experimental data. The prediction is then performed based on how similar the specific case is to the selected effect definition.

Trained classifiers also derive rules or patterns for their prediction task. However, this happens as part of the training of the corresponding classifier and is only influenced by the presented data samples. These machine learning devices use supervised learning (*i.e.* data samples are labeled; for instance as 'deleterious' or 'neutral') to generalize rules or patterns from the feature values present in the data. They require both positive and negative samples and operate under the assumption that the training data is representative of the prediction task. Popular examples of machine learning devices are support vector machines (SVM), multilayer perceptrons (MLP; also called artificial neural networks, ANN), decision trees (DT), random forests (RF), and rule-based learners. Major differences between these lie in the types of input features (*e.g.* numeric, ordinal, nominal) that can be handled and the way the final prediction is calculated. In terms of result interpretability we distinguish between black-box and white-box predictors. Black-box predictors (*e.g.* SVM, MLP, RF) offer no further information on how the result was generated, in the sense that the classification reasons are hidden in the

model and thus not interpretable by humans. White-Box predictors (DT, Rule-based learners) on the other hand can be interpreted by looking into the model and following the decisions or rules.

The data for training can be retrieved from databases such as the ones mentioned in section 1.3. Pathogenicity predictors are typically trained or optimized on variants with probable disease association as their task is to distinguish between natural variation and disease-causing mutations. This implies (i) disregarding variants that have molecular effects as long as these are not associated with diseases and (ii) a focus on human variants, as these are the ones for which disease annotations are predominantly available and for which predictions are most wanted. A 'neutral' prediction from a pathogenicity predictor does not indicate that the variant has no effect but rather that any possible effects are likely not involved in pathogenic phenotypes. Functional effect predictors are typically trained/optimized on variants with experimentally known molecular effects, which are predominantly obtained from non-human sequence experiments, as these are easiest to conduct. A positive prediction from a functional effect predictor does neither indicate nor exclude a possible human disease-association. It is thus difficult to compare methods with a different focus.

1.5 Thesis motivation and goals

Currently, there are significantly more disease variant predictors than effect variant predictors, which is attributable to the fact that medical research prudently focusses on revealing and targeting variants with disease association. However, there is also a need for accurate prediction of molecular variant effects. For instance also plant geneticists make use of computational methods to aid their research. Another question that can hardly be answered by disease prediction is how the human genome evolved functionally. In other words: what are the functional genetic differences that distinguish humans and apes? Understanding how variation affects phenotypes like disease onset

and progression requires us to understand how variants affect the underlying molecular processes.

In this work we focussed on improving the prediction of functional effects both in terms of accuracy and in terms of throughput over existing methods. We developed SNAP2, a classifier that outperformed current state-of-the-art methods and allows to predict every possible amino acid substitution in a protein in a time-efficient manner. Furthermore we visualized these predictions in heatmaps. This provides a simple but comprehensive representation that allows for intuitive interpretation of results and easy generation of hypotheses.

2 Methods

2.1 Data

Similar to an earlier publication (Bromberg and Rost, 2007), we used a mixture of experimentally determined neutral and effect variants from the Protein Mutant Database (PMD) and a set of putative neutral variants derived using the Enzyme Commission number (EC; Webb, 1992). We also added several putative effect variants from the disease-association databases OMIM and HUMSAVAR. In the following, the individual data components and their extraction will be briefly described:

The PMD provides functional variant annotations retrieved from literature reports and categorizes effects in seven classes with respect to the native protein function. We collected all variants and assigned them to either the 'neutral' or the 'effect' class in our data set depending on the reported effect. If the variant protein function was reported to be slightly ('-'), moderately ('--'), or substantially ('---') decreased, or if it was reported to be slightly ('+'), moderately ('++'), or substantially ('+++') increased, we labeled the variant as 'effect' variant. Only if the variant was annotated to cause 'no

change' ('=') in protein function, we assigned it to the 'neutral' class. In case of conflicting annotations (*e.g.* variant is reported as 'no change' with respect to one assay but as 'slightly increased' to another) the variant was assigned to the 'effect' class. This procedure yielded 38,179 effect variants and 13,638 neutral variants (*i.e.* a total of 51,817 variants) in 4,061 distinct proteins.

From the Online Mendelian Inheritance In Men (OMIM) and the HUMSAVAR we extracted variants associated with heritable diseases. Although, in many cases, there is no experimental evidence of their molecular functional effects, we assume these variants to have an effect on protein function. This is clearly an over-estimate, as these variants may just be in linkage disequilibrium with the causative variant or simply affect regulatory elements or splice sites. Thus, some of these variants may not have any effects on protein function. However, it is likely that this set is highly enriched in functional effect variants as they have been shown to exhibit a much stronger functional effect signal on average than the experimentally verified PMD variants (Schaefer et al., 2012). We thus collected 22,858 human variants in 3,537 proteins and assigned these to the 'effect' class.

From the above extraction steps we collected only 13,638 neutral variants as compared to 61,037 effect variants. This imbalance of available experimental 'neutral' and 'effect' variants suggests there is a significant selection bias towards experimental verification of effect variants. Researchers prudently focus on variants for which the strongest effects are expected, often with the goal to investigate certain diseases. Moreover, a detected effect variant is much easier to publish than a variant for which no functional effects could be measured. However, machine learning typically performs best when trained on a data set that is representative of the prediction task. While the true ratio of neutral and effect mutations in nature remains unknown, we chose to increase the fraction of neutral samples to obtain a close to balanced ratio.

We thus extracted putative neutral variants based on the following as-

sumption: If two independent experiments reveal that two similar, related proteins have the same enzymatic function we assume that most differences between these two are neutral with respect to this enzymatic function. While this approach explicitly neglects combinatorial effects such as compensatory mutations or effects on other possible functions, it is likely that this approach yields a set that is highly enriched in 'neutral' variants. We thus extracted all enzymes with experimental EC numbers from Swiss-Prot (UniProt Consortium, 2015) and did a pairwise alignment (using PSI-BLAST; Altschul et al., 1997) of all enzymes with the same EC number. Two more restrictions were imposed before considering any differences as neutral substitutions: (i) The sequences had to be $> 40\%$ identical and (ii) have HSSP-values > 0 (Sander and Schneider, 1991; Rost, 1999). Through this approach we extracted 26,840 variants in 2,146 proteins and assigned them to the 'neutral' class.

These three data sets were combined into our comprehensive training set. Thus the final data set consisted of 101,515 variants (40,478 neutral and 61,037 effect) in 9,744 distinct proteins (Hecht et al., 2015). Additionally, we used the 4,041 variants from the *E. coli* LacI repressor (Markiewicz et al., 1994) and the 336 variants from the HIV-1 protease (Loeb et al., 1989) as independent testing sets.

2.2 Features for variant effect prediction

The ability to predict variant effects through machine learning depends on the available information. For the task at hand, these features describe certain properties of proteins and amino acids. As mentioned before (Section 1.4), evolutionary information is a very important and thus commonly used feature. It provides information on the extent to which certain amino acids are observed in other species and other related sequences and thus can be used to estimate how likely an amino acid substitution is tolerated. Also the aforementioned biophysical amino acid properties are a commonly used feature as

these may give insights into the structural impact of a substitution. The feature calculation step involves extracting such information and transforming them into an appropriate machine-readable format. For instance, information on amino acid size may be given as a numeric value. If this is to be used, it has to be normalized in order to be used as an input value, because otherwise features with inherently large values would have proportionally larger impact on the prediction than features with inherently small values. Alternatively, that same size value can be transformed into binary or nominal inputs by defining thresholds (*e.g.* $small < n$, $n \leqq medium < m$, $large \geqq m$). The exact representation of a feature depends on the machine learning device and the available information. We extracted information from a variety of both experimental and predicted sources and calculated normalized numeric features where possible and multiple binary features otherwise. As the exact calculation is described in Hecht et al., 2015, the features used in the development of this method will only be briefly presented here.

We extracted evolutionary information from the PSI-BLAST position specific scoring matrix (PSSM: Altschul et al., 1997), from the position-specific independent counts profile (PSIC: Sunyaev et al., 1999) and by estimating co-evolving residues (Fodor and Aldrich, 2004; Kowarsch et al., 2010). Amino acid properties were retrieved from the AAindex database (Kawashima and Kanehisa, 2000). We considered biophysical properties (*e.g.* mass, volume, hydrophobicity, charge), structural propensities (*e.g.* c-beta branching, helix-breaker) and statistical properties (*e.g.* relative mutability, average flexibility, distance-dependent contact potentials). Structural variant information was not included through the use of known 3D structures as these are not available for most proteins. Instead we used predicted structural information such as secondary structure and solvent accessibility (Rost and Sander, 1993, 1994), as well as predicted residue flexibility (Schlessinger et al., 2006). Information on functional importance of residues was included in multiple ways. We used predicted protein-protein and protein-DNA interaction sites

(Ofran and Rost, 2007; Ofran et al., 2007), experimental annotation from Swiss-Prot if available, and whether a residue and its surroundings match any Pfam (Punta et al., 2012) or PROSITE (Sigrist et al., 2010) pattern. Moreover, we included a set of global sequence properties such as the amino acid composition, the secondary structure and solvent accessibility composition, the protein length, and low-complexity regions.

Where applicable we also used deltas of these features. These delta-features were calculated by comparing feature values for wildtype and mutant amino acid and estimating the strength and direction of the change attributable to the substitution. We also considered the immediate sequence environment for each substitution by including not only the variant residue position but also surrounding positions. Thus, we extracted each feature in a window between 3 and 21 residues (*i.e.* the central variant position and each 1-10 residues up- and downstream).

2.3 Prediction method

Machine learning offers a variety of methods for learning patterns from labeled data and using these to predict labels for novel samples. In order to identify the most suitable tool for our problem we initially trained and tested several tools from the WEKA suite (Frank et al., 2004) on our data. We used support vector machines, neural networks, decision trees, and random forests with default parameters and compared their performance. On our data neural networks and SVMs performed slightly better than decision trees and random forests. The difference between SVMs and neural networks was not significant. We proceeded with standard feed-forward neural networks because of better runtime and memory efficiency when applied through the Fast Artificial Neural Network (FANN: Nissen, 2003) library. The networks were designed to have two output nodes, one for 'neutral' and one for 'effect'. The following sections describe the method development and previously established training procedures (Hecht, 2011).

2.3.1 Clustering and cross-validation

In order to avoid over-fitting of the model and over-optimistic performance estimates we created 10 sub-sets from our data in such a way that there was no significant sequence similarity between any two proteins in different sub-sets. We first used PSI-BLAST for each sequence against our entire set and collected all significant hits ($E - value < 10^{-3}$) for each sequence. We then built an undirected graph where each protein was assigned to a vertex and connected through edges with all vertices for which we had recorded a significant hit. We then clustered our data through single-linkage clustering and thus collected all proteins into the same set that were reachable through a path in a graph. This approach yielded 1,241 clusters with the number of members ranging from 1 to 1,941. From this clustering we created ten sub-sets in such a way that we assigned all proteins and all their variants from the same cluster to one of ten sub-sets while balancing (as far as possible without separating cluster members) the overall number of variants per set as well as the ratio of 'neutral' and 'effect' variants within each set.

These ten sub-sets were used in the cross-validation of our method by using eight sets for training, one set for optimization and one set for final testing. We cycled through these set combinations such that each set was used exactly once for optimizing and exactly once for testing in different combinations. This ensured that no variant from the same or any similar protein was ever used simultaneously for training and testing. During development of the method, we always used eight sets for training and one optimization set for testing. The respective tenth final testing set was only used after development to reliably estimate the final network performance as reported in section 3.

This approach effectively yielded ten different networks, each optimized and tested on a different sub-set of our data. In order to avoid the risk of over-fitting we did not select the best of the ten networks for the final method, but instead decided to use all ten networks. The final prediction

is calculated as a jury-decision by using the average 'neutral' output and the average 'effect' output over all ten networks. From these two scores we calculated a single output score as the difference between the average 'effect' output and the average 'neutral' output which resulted in a score ranging from -1 (all networks predict 100% neutral probability) to +1 (all networks predict 100% effect probability). For simplicity, we re-normalized this score to integers between -100 and +100 in the final method output.

2.3.2 Feature selection and parameter optimization

Determining the optimal feature combination for a prediction task is highly non-trivial for several reasons: First of all, we do not know for certain to which extent protein and variant properties are relevant for the task, or even if we know all the relevant factors. Moreover, the corresponding importance of each feature varies between different samples (*e.g.* variant prediction in membrane proteins has other priorities than in globular proteins). As the 'optimal' combination depends on the prediction task, it is subject to how representative the training data is for this task. We developed a general functional effect predictor by generalizing patterns from different kinds of proteins, because we lack the data to develop a specialized method for every problem. This meant sifting through a large space of potentially relevant features in order to identify those that perform best on this general task. Moreover, finding the optimal combination is also difficult because it would involve testing every possible combination for all features, which is possible but extremely CPU time-consuming. Instead, we used heuristic methods that are likely to find at least a good combination, while cutting CPU requirements to a fraction of the exhaustive search.

As mentioned above (Section 2.3.1) we separately trained and optimized ten networks. For each of these, we individually selected the best feature combination by standard greedy forward selection. The following steps were applied for all ten networks separately: First, we trained each network on

all features individually and estimated their individual performance on the cross-training set by calculating the area under the receiver operating characteristic curve (Section 2.4). We then selected the best-performing feature, paired it with all others and selected the best-performing duo. This procedure was repeated, always adding the best-performing feature of each round to the previous best-performing combination until no additional feature yielded an improvement. In the end we tested the best feature combination on the testing set (the tenth set that had not been used during feature selection) to assure that the network was not overfitted. A combined feature set was obtained by collecting all features that improved performance on any of the ten networks into one feature set. We trained all networks using this feature set and then performed a backward elimination selection by which we removed all features that could safely be removed without lowering the overall performance. The remaining features constituted our final feature set.

The final feature set was used to find the optimal network architecture. For each of the ten networks we applied heuristically selected parameter combinations: learning rate between 0.005-0.1, learning momentum 0.01-0.3, and hidden units 10-100. We again tested each parameter combination separately for each network using the cross-training set and selected the combination that performed best. In the end, the performance on the cross-training sets was compared to the performance of the corresponding testing sets to assure that there was no significant overfitting.

2.3.3 Alignment-free prediction

Of all features, alignments of related protein sequences carry the most information on whether or not a variable is acceptable. As many (approximately 10%-20%) sequences in today's databases continue to not map to any known sequence, there is no evolutionary information available for these so-called orphan sequences. In fact, even for human there are currently (Feb 2015) over 600 sequences for which less than 5 homologs can be found. We thus

specifically trained a method to perform well when no evolutionary information is available. Towards this end, we created a substitution score matrix from the predictions made by our regular method and used this matrix as an independent feature along with all other features that did not require alignments. Networks were trained and optimized as above. The resulting prediction method constitutes a fall-back mode for the main method. Although by definition weaker than the main method, this enabled predictions for cases that would otherwise not be predictable.

2.4 Performance measures

We employed several different performance measures to evaluate different aspects of method performance. The following standard definitions were used: True positives (TPs) were correctly predicted variants with experimentally annotated effect. True negatives (TNs) were correctly classified variants for which experiments confirmed no effect (*i.e.* neutrals). And correspondingly, false positives (FPs) and false negatives (TNs) were incorrectly classified samples with or without experimentally annotated effect, respectively. We considered three levels of performance: the inner-class performance, the combined class performance and the overall model performance.

The inner-class performance was estimated using the standard formulas for precision and recall (*i.e.* accuracy and coverage, respectively) for both the 'effect' and the 'neutral' class separately:

$$Precision_{effect} = Accuracy_{effect} = \frac{TP}{TP + FP}, \tag{1}$$

$$Recall_{effect} = Coverage_{effect} = TPR = \frac{TP}{TP + FN}, \tag{2}$$

$$Precision_{neutral} = Accuracy_{neutral} = \frac{TN}{TN + FN}, \tag{3}$$

$$Recall_{neutral} = Coverage_{neutral} = \frac{TN}{TN + FP}. \tag{4}$$

The recall value of the effect class also represents the true positive rate (TPR; Eqn. 2). Combined class performances were measured through the F1 measure in order to comprehensively estimate the performance for neutral and effect variants individually:

$$F1_{effect} = \frac{Precision_{effect} * Recall_{effect}}{Precision_{effect} + Recall_{effect}}, \tag{5}$$

$$F1_{neutral} = \frac{Precision_{neutral} * Recall_{neutral}}{Precision_{neutral} + Recall_{neutral}}. \tag{6}$$

The overall performance was used to compare different models and methods. We therefor calculated the Matthews correlation coefficient (MCC) and the overall two-state accuracy (Q2):

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}, \tag{7}$$

$$Q2 = \frac{TP + TN}{TP + FP + TN + FN}. \tag{8}$$

The false positive rate (FPR) was calculated as

$$FPR = \frac{FP}{FP + TN}. \tag{9}$$

Moreover the area under the curve (AUC) of the receiver operating characteristic (ROC) was used for model comparisons. This curve was obtained through plotting the true positive rate (Eqn. 2) against the false positive rate (Eqn. 9) at all possible decision thresholds (*i.e.* every score from -100 to +100).

We estimated standard error and standard deviation by bootstrapping. The 1000 bootstrap sets were created by randomly selecting 50% of all variants without replacement. Although this is typically done with replacement, we experienced that bootstrapping without replacement yields more accurate estimates due to overrepresentation of certain protein families. With these n=1000 sets we calculated the standard deviation (StdDev; Eqn. 10) as the average performance difference of each set (x_i) from the overall performance average (χ). The standard error (StdErr; Eqn. 11) was calculated by dividing the standard deviation by the square root of sets:

$$StdDev = \sqrt{\sum (x_i - \chi)^2}, \tag{10}$$

$$StdErr = \frac{StdDev}{\sqrt{n-1}}. \tag{11}$$

2.5 Result visualization

One important achievement of this method is its ability to efficiently predict large amounts of mutations in a protein. This allowed us to expand the standard resolution from predicting single mutations to predicting the entire mutability landscape of a protein (Hecht et al., 2013). For an average-sized protein (*i.e.* roughly 300 residues) we can predict all non-native substitutions at every position in less than one hour of runtime, yielding a total of roughly 5,700 predictions per protein. However, this much information cannot be reasonably represented in text form and thus requires appropriate visualization. We therefore implemented a JAVA script component with the ability to represent this data as a heat map (Yachdav et al., 2014). This allowed showing all 19 substitutions (y-axis) for every residue of the protein (x-axis) along with their predicted effect by color-coding the numerical prediction in a color cascade (Fig. 7b,c) ranging from green (Score -100, most reliable neutral prediction) over white (Score 0, inconclusive prediction) to

red (Score +100, most reliable effect prediction). This component was included in a web-server (available at http://rostlab.org/services/snap2web) in order to enable users to view and download both the numerical and the visual output of our method.

3 Results and discussion

During the course of this thesis we developed SNAP2, a neural network based classifier for the prediction of functional effects of single amino acid substitutions in proteins. It is able to accurately predict variants even in orphan proteins and presents prediction results for all possible substitutions in comprehensive heat maps. The following sections present comparisons to other state-of-the-art methods and describe the improvements made in our novel method. We show the heat map representation along with a possible use-case and discuss difficulties that users may face when interpreting results from variant effect predictors.

3.1 Better prediction of functional effects

For the final method we estimated an overall performance of over 83% two-state accuracy. These estimates were obtained from our cross-validation testing sets, that had no part in method development. We first compared our new method SNAP2 against it's predecessor SNAP, as both methods shared a significant amount of training data. Towards this end, we assessed the relative benefit of the newly implemented features by training our method on exactly the same data that SNAP was trained on. Through this approach we observed that the additional features accounted for approximately 1.2% performance increase over the original SNAP (79.8±0.4% up to 81±0.5% Q2; Eqn. 8). This performance increase was significant, although the original SNAP had an advantage in this comparison as it had been trained on the

same data, while the SNAP2 estimates were taken from the cross-validation (and had thus not been used in training). Next we assessed the benefit of adding disease-associated variant data to our training data. We trained our method on the extended training set but tested against the original SNAP data set as before. Here we observed an even higher performance increase: SNAP2 (82.4±0.4% Q2) outperformed its predecessor SNAP (79.8±0.4% Q2) even more, which suggested that the additional data had indeed improved performance. The improvement is also visible in figure 2. Next, we compared our methods against two other state-of-the-art functional effect predictors: SIFT (Kumar et al., 2009) and PolyPhen-2 (Adzhubei et al., 2010). While there are many methods for variant effect prediction (Section 1.4), it would have been inappropriate to test disease-association predictors on our data. We therefore explicitly only included these two methods because they had been optimized for functional effect prediction according to the authors. Figure 2 shows that SNAP2 (again, predictions were taken from cross-validation sets that were not used for training) compares favorably to other methods.

PolyPhen-2 has been explicitly optimized on human variants. Although the authors claimed that their method should work equally well on other eukaryotes, we felt that the above comparison put PolyPhen-2 at a disadvantage as only 25% of our data consisted of human variants. We therefore re-calculated performance values separately for human and non-human PMD variants (Sections 1.3 and 2.1). As can be seen in Table 1, PolyPhen-2 does indeed perform significantly better on human variants on which it performs on par with SNAP2, although PolyPhen-2 values may be overestimated due to substantial overlap with its training data. Both methods exhibit better performance than SNAP and SIFT. On non-human variants however, SNAP2 significantly outperformed all other methods in terms of $F1_{effect}$(Eqn. 5), MCC (Eqn. 7) and Q2 (Eqn. 8).

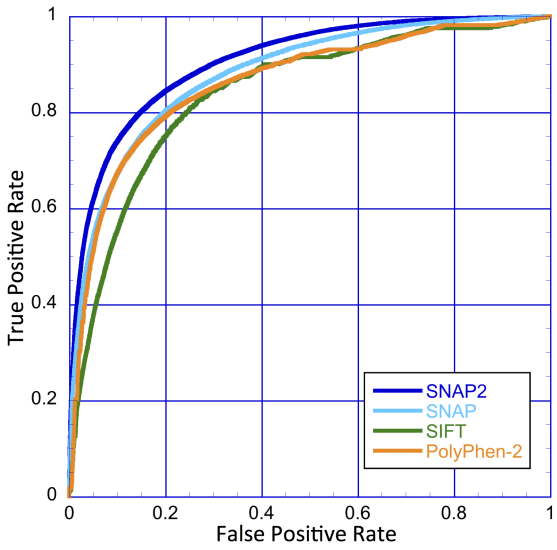

Figure 2: **Receiver Operating Characteristic (ROC) comparison.** Shown are the receiver operating characteristics for our new method SNAP2 (dark blue, AUC=0.905), it's predecessor SNAP (light blue, AUC=0.880) and the two widely used functional effect predictors SIFT (green, AUC=0.838) and PolyPhen-2 (orange, AUC=0.853). Curves are significantly different ($P < 10^{-4}$) according to the method by DeLong et al., 1988. (Figure adapted from Hecht et al., 2015)

	Method	$F1_{effect}$	$F1_{neutral}$	MCC	Q2
human	SNAP2	**78.0% ± 0.6**	46.3% ± 1.3	0.24 ± 0.01	**68.8% ± 0.7**
	PolyPhen-2	**78.4% ± 0.4***	45.1% ± 1.1*	0.23 ± 0.01*	**68.9% ± 0.5***
	SNAP	74.9% ± 0.5*	46.7% ± 1.1*	0.22 ± 0.01*	65.8% ± 0.6*
	SIFT	72.2% ± 0.6	**49.0% ± 1.0**	0.23 ± 0.01	63.6% ± 0.6
non-human	SNAP2	**79.9% ± 0.3**	45.8% ± 0.8	**0.26 ± 0.01**	**70.7% ± 0.4**
	PolyPhen-2	77.1% ± 0.4	44.7% ± 0.8	0.22 ± 0.01	67.6% ± 0.5
	SNAP	77.2% ± 0.3*	45.5% ± 0.9*	0.23 ± 0.01*	67.9% ± 0.5*
	SIFT	77.0% ± 0.3	45.8% ± 0.8	0.23 ± 0.01	67.7% ± 0.4

Table 1: **Method performances on PMD data.** For each method, SNAP2, PolyPhen-2, SNAP and SIFT, the corresponding performance is shown separately for human and non-human proteins. Performance values were calculated for $F1_{effect}$ (Eqn. 5), $F1_{neutral}$ (Eqn.6), MCC (Eqn. 7) and Q2 (Eqn. 8) measures. The data consisted of 9,657 human variants in 678 proteins and 42,160 variants in 3,383 non-human proteins. Significantly best results for each measure are highlighted in bold. Marked values (*) indicate potentially over-estimated performance due to substantial overlap with training data. SNAP2 values were taken from cross-validation sets that had not been used for training.

3.2 Difficult cases and method combinations

When comparing different prediction methods, one will notice that overall method performance is quite similar. However, predictions can differ substantially for individual cases. For instance, Liu et al., 2011 found that the pairwise agreement between any two methods lies between 61% and 77%. To turn this observation into a concept we grouped the variants that could be predicted by all four predictors (83,671 variants) into three classes. We found 53,976 'easy' cases (every method gets them right), 23,630 'difficult' cases (at least one but not all of the tested methods gave a correct prediction), and 6,066 'unsolvable' cases (no method gave a correct prediction). As can be seen from these numbers, the bulk of method performance is gained from easy cases. It appears that a well-trained method will achieve a minimum of 68% accuracy by simply using the most informative features, that almost every current implementation employs. It is likely that most of these variants can be predicted by simply looking at the alignment of related proteins. For unsolvable cases, on the other hand, it might be that alignment information is either not sufficiently available or misleading.

We thus compared performances on those difficult cases (Figure 3) and found that our new method significantly outperformed the other methods with 67.2% accuracy. SNAP and PolyPhen-2 performed similarly with 55.2% and 57.8% accuracy respectively, although PolyPhen-2 was at a disadvantage (as it is specifically optimized for human variants). SIFT (45.5%) performed within the standard deviation of the random prediction model (44.5%), which assumed a background of 60% effect and 40% neutral like the overall data set (*i.e.* it randomly predicted effect variants slightly more often than neutral variants).

We investigated the properties of these cases by looking into the human cases that SNAP2 could correctly predict while the others could not. We found that these were located at positions at which the variant amino acid was observed in the alignment (and in some cases even more frequently than

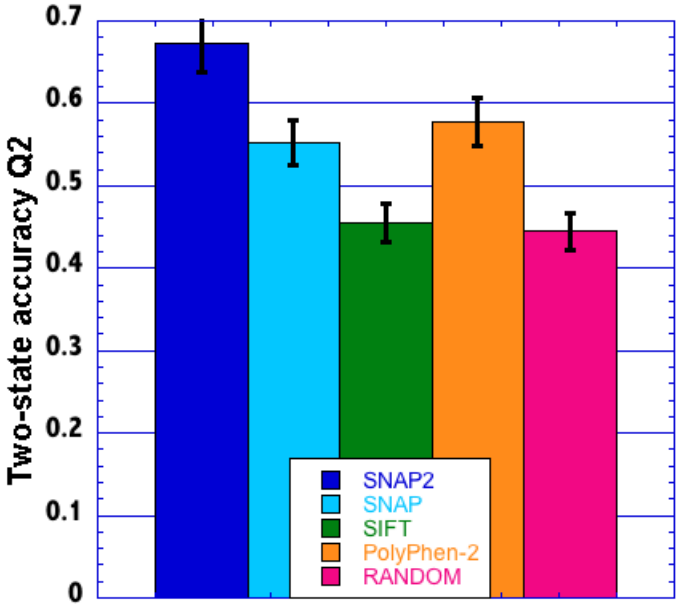

Figure 3: **Comparison on difficult cases.** The average two-state accuracy (Q2; Eqn. 8) on all difficult cases is shown for SNAP2 (dark blue), SNAP (light blue), SIFT (green) and PolyPhen-2 (orange). For reference the random prediction model with a 60:40 effect:neutral background is shown in pink.

the human native amino acid). As mentioned above, the presence of an amino acid in the alignment is a strong indicator that such a variant can be tolerated, thus possibly misleading other methods to predict these as neutral. This suspicion was further supported by the fact that our alignment-free version of SNAP2 predicted 75% of the effect variants with over 90% accuracy - a value that is significantly higher than the average performance of alignment-free SNAP2. This indicated that for these cases SNAP2 greatly profited from the variety of additional features rather than over-relying on alignment information.

Can the reliability of prediction be increased by using multiple methods? We investigated the benefit of combining prediction methods for the human variants of our data by employing a combination of SIFT and PolyPhen-2 that only considered predictions if both methods agreed in their prediction. This naïve combination was compared to our new method SNAP2 and to each of the individual methods (SIFT and PolyPhen-2). Figure 4 shows that the method combination performed slightly better than SIFT on the neutral cases but did not improve over any of the individual methods on the effect variants. Moreover, the combination did not perform any better than SNAP2 throughout the curves. This suggests that users should not blindly combine methods and trust the results if methods agree but rather choose a method specifically for their problem or employ specifically optimized method combinations such as PredictSNP (Bendl et al., 2014) or Condel (González-Pérez and López-Bigas, 2011).

3.3 Neutral variant dilemma

As can be seen from Table 1 and Figure 4, neutral variants from our PMD set were predicted much worse than effect variants by all methods. In accordance with the findings reported by Bromberg et al., 2013, this can be attributed to a bias in both variant selection and experimental verification. The variant selection bias results from the fact that variants are not investigated

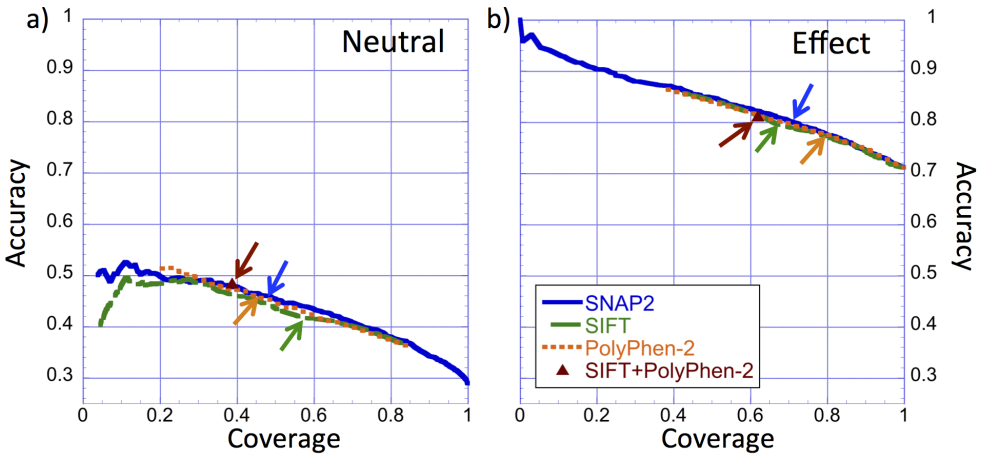

Figure 4: **Naïve method combination is not better than individual methods.** This figure shows accuracy versus coverage separately for neutral (Panel a) and effect (Panel b) variants. The accuracy curve is the percentage of correctly predicted neutral (Eqn. 3, panel a) or effect (Eqn. 1, panel b) variants out of all predictions at the given threshold. Correspondingly the coverage is the percentage of correctly predicted neutral (Eqn. 4, panel a) or effect (Eqn. 2, panel b) variants out of all observations at the given threshold. The default thresholds of each method are marked by arrows for SNAP2 (dark blue), SIFT (green), PolyPhen-2 (orange) and the naïve combination of SIFT and PolyPhen-2 (brown). (Figure adapted from Hecht et al., 2015)

randomly. Researchers focus on variants for which the most extreme effects are expected or on variants that are hypothesized to be involved in diseases, which explains the imbalance in numbers between neutral and effect variants in today's databases. This poses a significant obstacle for computational approaches, which require sufficient amounts of data in all classes in order to generalize. The lack of experimentally verified neutral variants hampers the ability to extract relevant patterns for neutral variants and lowers the predictive performance. The incomplete verification bias, on the other hand, results from the triviality that 'not observed' does not necessarily imply 'not existing'. Negative experiments are much harder to carry out as they would require to assay every possible effect. Variants are thus typically evaluated on the basis of one or a few assays. If no effect is observed the variant is reported as neutral with respect to that assay but it may well have an effect on a different phenotype/assay. As today's data is too scarce to develop specific methods for each phenotype, we have to accept that our general approach is skewed with respect to neutral variants. This may have effects on the extraction of patterns for neutral variants as well as on the evaluation of performance.

3.4 Interpretation and reliability of variant prediction

A major pitfall in the application of variant effect prediction for biologists and geneticists is the choice of method and the interpretation of results. There are significant differences between the seemingly similar approaches of functional effect and disease-association predictors. Disease association methods predict the likelihood of variants to be involved in diseases by distinguishing disease variants from the background of natural genetic variation. For many users 'disease-associated' implies that a variant must have strong functional effects. This is however only true for the simplest cases. For instance, in some monogenic or Mendelian diseases, the molecular effect can be quantified and directly linked to the phenotype (Hamosh et al., 2005).

Yet, most disorders appear to be complex in the sense that they involve multiple genes, variations and/or environmental conditions, as was shown by GWAS (Wellcome Trust Case Control Consortium, 2007; McCarthy et al., 2008). It has also been shown (1000 Genomes Project Consortium et al., 2010) that disease-associated variants can be found in healthy individuals. Moreover, even in seemingly clear cases the definition of a disease variant can be difficult. For example, certain variants of the hemoglobin B-chain cause a condition called sickle-cell anemia, which is associated with chronic health issues. On the other hand, that same condition grants immunity to some types of Malaria. In other words: the definition of a disease variant can depend on the environment and genotype of the individual, making the interpretation of 'disease'-predictions difficult.

A different set of problems applies for functional effect prediction. In contrast to disease prediction, functional effect prediction focusses on the native molecular protein function. These predicted effects are independent of the individual and often also independent of the environment but offer no simple interpretation of their biological relevance. For instance, a weakly predicted effect in p53 may have a massive biological impact on the individual whereas a strongly predicted effect in another protein may have little biological significance. Moreover, functional effect predictors cannot distinguish the direction of predicted effects. That is, they cannot discern between gain and loss of function, but simply predict that a variant has an effect on the native protein function. This means that predictions have to be interpreted with respect to the protein under investigation.

Today's methods can distinguish between sets of variants that are highly enriched in effects and sets that are not (Schaefer et al., 2012), but pinpointing the one mutation that causes a certain phenotype within an individual genome is often beyond our reach. In order to aid the prioritization of experimental variant testing, users need to be able to focus on the most promising candidates. Towards this end, we provided a reliability index (RI) that allows

users to zoom in on the most reliable predictions. We first fixed the threshold for the binary categorization of our prediction by selecting a threshold on our balanced data set (Fig. 5a) that provided the highest overall performance for both classes. We then calculated reliability bins from the output score by projecting the score onto integers between 0 and 9. From these bins we estimated the performance of each reliability index for our training data, thus providing users with an estimate of accuracy for each index. Figure 5b depicts the cumulative accuracy and coverage that can be expected above each reliability threshold and to how many samples of our data this applied. For example, the grey arrows mark predictions made at reliability index of 7 or above, which applied to over 58% of our data. For this most strongly predicted half of our data, we estimated over 90% accuracy for neutral samples and almost 95% for effect variants. Focussing on these variants with RI 7 or better can thus be considered a reasonable approach towards prioritizing promising candidates for experimental verification.

3.5 Variant prediction in orphan proteins

Orphan proteins are sequences that find no (or, by extension, only very few) related sequences in today's databases. In other words, there is no meaningful alignment available for the prediction of variant effects. Most methods today rely heavily on the evolutionary information encoded in alignments and thus perform poorly at best for these cases. Ongoing sequencing efforts continue to bring in novel sequences and reveal novel protein families. In fact, the number of orphan families keeps increasing. In October 2012 the UniRef50 consisted of roughly 5.5 million sequence clusters of which over 3.5 million contained only one sequence, meaning that approximately 64% of all known protein clusters (clustered at 50% sequences identity) were represented only by a single orphan protein. Over the last 2 years, this number has more than doubled. The current release of UniRef50 (Feb 2015) lists over 13.2 million sequence clusters with roughly 8.2 million consisting of a single sequence. For

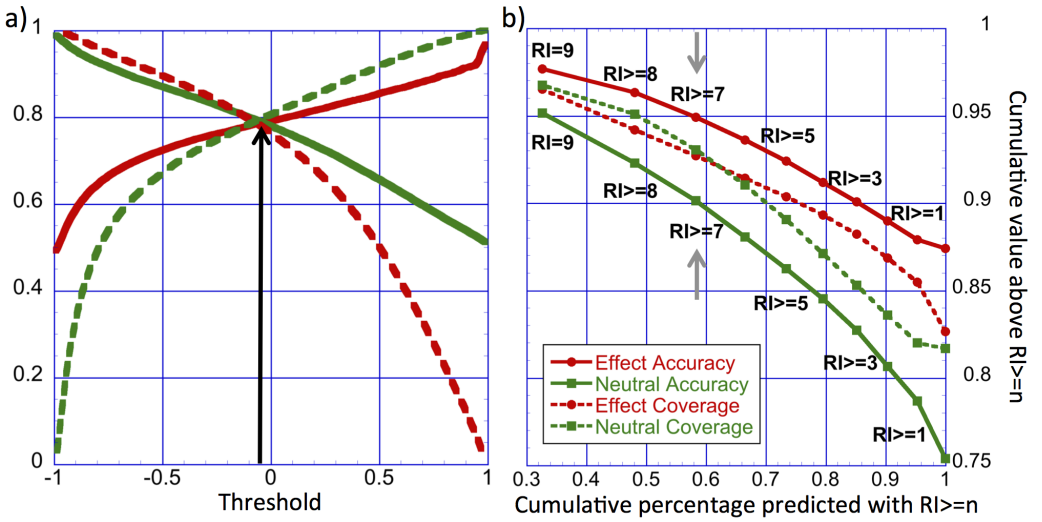

Figure 5: **Threshold and reliability index.** Panel (a) shows accuracy (solid lines, Eqn. 1 and 3) and coverage (dashed lines, Eqn. 2 and 4) for both effect (red) and neutral (green) variants over the entire spectrum of possible thresholds. A black arrow marks the threshold selected for the binary prediction. Panel (b) depicts the same measures above certain reliability indices (RI). The leftmost point (RI=9) corresponds to predictions with the highest reliability, while the rightmost point (RI≥0) includes all predictions. Shown are the cumulative accuracy (solid lines) and cumulative coverage (dashed lines) above the corresponding reliability value (ranging from 0, lowest reliability to 9, highest reliability) separately for effect (red) and neutral (green) variants. Grey arrows mark the reliability index that applies to the most strongly predicted half of the tested data. (Figure adapted from Hecht et al., 2015)

these 8.2 million proteins/families no other sequences can be found that are more than 50% identical. While sequences can be quite similar in structure and function with less than 50% sequence identity, this still suggests that many proteins in today's databases are orphans or will have very sparse and thus possibly uninformative alignments.

We specifically aimed at predicting variants for these proteins by training a method that did not require alignments (Section 2.3.3). SNAP2$_{noali}$ was first tested on our entire data set through cross-validation. With an overall accuracy of Q2=68% it performed significantly worse than other methods on our entire data set, which was to be expected and highlighted the importance of alignment information (Fig. 6).

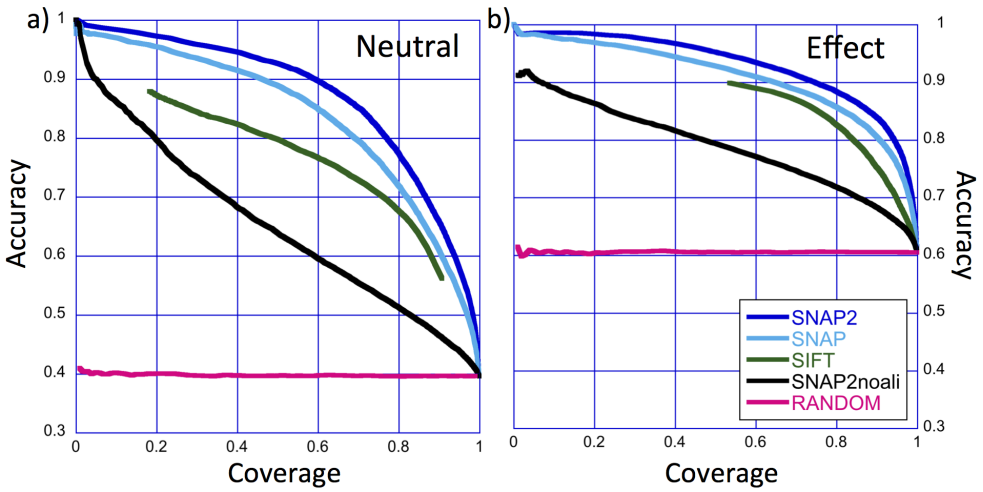

Figure 6: **Method performances on the entire training data.** Shown are the accuracy versus coverage curves for SNAP2 (dark blue), SNAP (light blue), SIFT (green), SNAP2$_{noali}$ (black) and the random (pink) prediction model with 60:40 effect:neutral background for neutral (panel a) and effect (panel b) variants.

To test the performance of our method on orphan proteins we sorted our data by the amount of sequences in the corresponding alignments. We found no real orphan (*i.e.* no protein for which no significant hit was found) in

our data. However, by raising the threshold to less than five hits, we identified a small number of 8 'orphan' proteins with a total of 248 variants. On these, our alignment-free method SNAP2$_{noali}$ achieved an accuracy of 62%, while our regular method SNAP2 performed at 61% accuracy. PolyPhen-2 predicted only three of the eight proteins (103 variants) at 60% and SIFT gave no predictions, which corresponds to random. Although this was a very small test set, the results indicate that the alignment-free method constitutes an important alternative for predicting variants in proteins with small alignments and is likely the best solution for orphan proteins.

3.6 The protein mutability landscape

The high computational power of today's processors and the improved efficiency of our algorithm allowed us to shift the predictive scope of our method. Instead of only predicting one mutation at a time, we expanded the view by sketching the entire mutability landscape of a protein. This landscape can be defined as the predicted effects of each non-native substitution at each position in the protein (Fig. 7b).

This comprehensive view of functional effects in proteins brings about opportunities but also a number of challenges. The human beta-2-adrenergic receptor (UniProt ID: ADRB2_HUMAN, UniProt AC: P07550) is an integral membrane protein with seven transmembrane helices crossing the lipid bilayer. These seven transmembrane helices are clearly visible from the predicted effects shown in Figure 7b and align with the DSSP-assigned (Kabsch and Sander, 1983) secondary structure elements (Fig. 7a) based on the high resolution structure 3PDS (Rosenbaum et al., 2011) in the Protein Data Bank (PDB; Bernstein et al., 1977). This could be expected, as transmembrane regions are often well conserved due to the specific biophysical and structural restrictions imposed upon them. Still, there is a remarkably high number of effect predictions even for amino acids that fulfill the transmembrane requirements (*i.e.* that are likely structurally acceptable/neutral) as

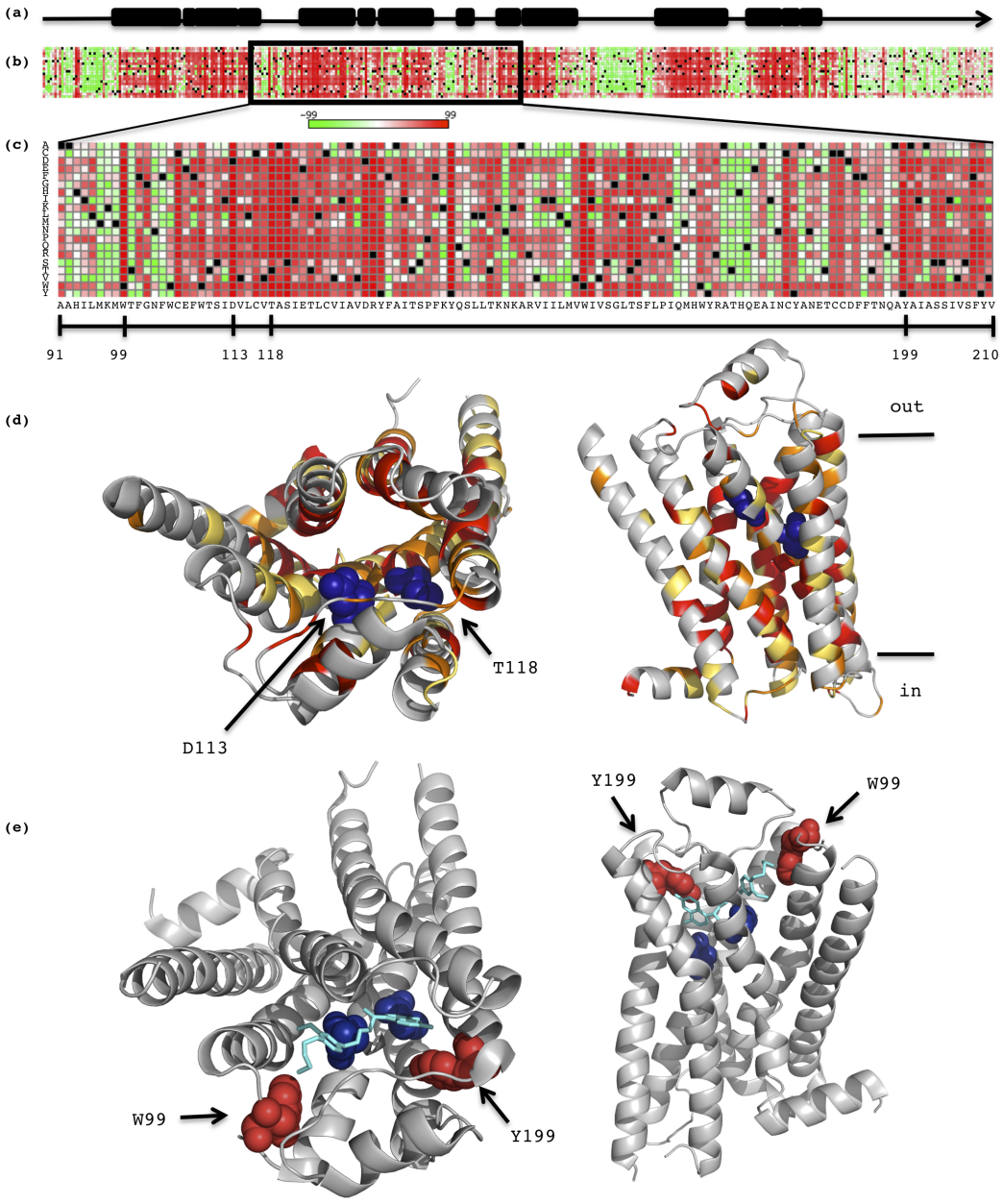

Figure 7: **Mutability landscape of the human beta-2-adrenergic receptor.** Figure caption on the next page.

Figure 7: **Mutability landscape of the human beta-2-adrenergic receptor.** Panel (a) shows secondary structure assignments from DSSP based on the protein structure PDB: 3PDS. The predicted mutability landscape is shown in panel (b) with the x-axis representing the entire sequence. Predicted effects (ranging from green/neutral over white/inconclusive to red/effect) are shown on the y-axis for all 19 non-native substitutions. Panel (c) zooms in on the region spanning from residue 91 to residue 210. It shows the two binding sites D113 and T118 along with the two strongly predicted and highly correlated residues W99 and Y199. Panel (d) shows the 57 predicted high effect residues on the structure colored according to their predicted effects. Grey represents no or little effect (SNAP score $\leq$ 20), yellow depicts low effects (20 $<$ score $\leq$ 40) , medium effects are marked in orange (40 $<$ score $\leq$ 60), and high effects are shown in red (score $>$ 60). Panel (e) shows the structure (PDB: 3PDS) of human beta-2-adrenergic receptor (UniProt ID: ADRB2_HUMAN; UniProt AC: P07550) with an irreversibly bound agonist (cyan sticks), the two known binding site residues D113 and T118 (blue spheres) and the two predicted high effect residues W99 and Y199 (red spheres) that exhibit strong residue couplings with each other and the binding site. (Figure adapted from Hecht et al., 2013)

well as for non-transmembrane variants. We found 57 residues that were predicted to cause strong effects upon mutation (average substitution effect score of over 60). To visualize these we used the available structure (PDB: 3PDS; Fig. 7d) and colored all residues according to their average predicted substitution effect. Figure 7d shows no or very small effects (SNAP score $\leq$ 20) in grey, strong effects (SNAP score $>$ 60) in red, medium effects (40 $<$ SNAP score $\leq$ 60) in orange, and weak effects (20 $<$ SNAP score $\leq$ 40) in yellow. Notably, the vast majority of effect predictions are facing the two known binding site residues (blue spheres) or are located in their proximity. Only few of these have been experimentally tested for functional effects, which makes an assessment difficult: among these 57 positions, we could only find 11 experimental effect annotations. What about the remaining 46 residues without experimental evidence? Some of these may simply be false positives or predicted to have an effect because of structural importance but it appears likely that some of these may also be directly relevant for function and their involvement is yet unknown. A similar observation had previously been made by Bromberg et al. (2009) in an *in-silico* mutagenesis study of the human melanocortin-4 receptor. Following their example, we filtered our results by only considering variants that could be considered neutral given their biophysical properties. Towards this end we used the PHAT matrix (Ng et al., 2000) for transmembrane regions and the BLOSUM (Henikoff and Henikoff, 1992) for all other regions and filtered out all predictions for variants with a negative substitution score in these matrices. This filtering left us with twelve predicted high-effect positions that are likely to directly affect function upon mutation. These twelve can already be considered reasonable candidates for experimental testing in this protein but other proteins may exhibit significantly more high-impact sites. In order to further narrow down candidates, we studied these variants in the light of correlated mutations by applying EVfold (Marks et al., 2012; Hopf et al., 2012). This tool was designed to predict inter-residue contacts from correlated sequence variation

and thus possibly predict the 3D structure. We looked at the correlation patterns for our twelve candidates and found that only two of these had residue couplings among the top 5%. W99 and Y199 (Fig. 7e, red spheres) were strongly correlated with each other and the two known binding site residues (Fig. 7e, dark blue spheres). A visual inspection of the protein structure (PDB: 3PDS) with irreversibly bound agonist (Fig. 7e, cyan sticks) also suggested functional importance. This example shows how the protein mutability landscape of functional effects can be used to generate hypotheses for drug development and variant prioritization.

4 Conclusion

This thesis presents a newly developed method for the prediction of functional effects of amino acid substitutions in proteins. The method is shown to perform better (on non-human variants) or on par (on human variants) with current state-of-the-art methods, while significantly outperforming all competitors on the difficult cases. The novel features, that no other method uses, allowed to make accurate predictions where other methods struggled and even allowed to make non-random predictions for orphan proteins. The method presented in this thesis is the most reasonable choice for all proteins with little or no evolutionary information and capable of making predictions where other methods cannot.

Better CPUs and an improved algorithm allowed us to shift the predictive focus of our method from single variants to entire mutability landscapes. These landscapes have been shown to be promising and may enable new insights into protein engineering, drug development and variant prioritization. They may be the key for studying the genotype-phenotype link on a molecular level if we learn how to correctly interpret them. While the experimental counterparts remain constrained by the substantial amount of resources required for mutagenesis studies, computational prediction is cur-

rently constrained by the amount of available experimental data. Comprehensive testing studies such as performed for the HIV protease or the LacI repressor are invaluable sources of information as they help overcome the neutrality dilemma and broaden our understanding of functional variant effects. Still, computational prediction of variant effects is crucial for coping with the ever-increasing amount of sequencing data and will likely play a major role in understanding the molecular mechanisms that link genetic variation and diseases.

Acknowledgements

I would like to thank Prof. Dr. Burkhard Rost for supervising my thesis, for giving me advice and criticism, and providing the high-end working environment to carry out my studies. Thanks to my colleagues at the Rostlab for interesting discussions and good times both in and outside of the lab. Special thanks to Timothy Karl and Laszlo Kajan for their patience in providing technical assistance and stable hard- and software. Thanks also to Marlena Drabik for all the little things that often (appear to) go unnoticed and her tireless efforts in dealing with administrative issues. Thanks also to Edda Kloppmann and Peter Hönigschmid for proof-reading the manuscript.

Last, not least, I would like to thank my friends, my family and especially my girlfriend Sarah who continuously supported me in any imaginable way.

References

1000 Genomes Project Consortium, Abecasis, G. R., Altshuler, D., Auton, A., Brooks, L. D., Durbin, R. M., Gibbs, R. A., Hurles, M. E., and McVean, G. A. (2010). A map of human genome variation from population-scale sequencing. *Nature*, 467(7319):1061–73.

1000 Genomes Project Consortium, Abecasis, G. R., Auton, A., Brooks, L. D., DePristo, M. A., Durbin, R. M., Handsaker, R. E., Kang, H. M., Marth, G. T., and McVean, G. A. (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422):56–65.

Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., Kondrashov, A. S., and Sunyaev, S. R. (2010). A method and server for predicting damaging missense mutations. *Nat Methods*, 7(4):248–9.

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997). Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic Acids Res*, 25(17):3389–402.

Barrett, J. C. and Cardon, L. R. (2006). Evaluating coverage of genome-wide association studies. *Nat Genet*, 38(6):659–62.

Beck, T., Hastings, R. K., Gollapudi, S., Free, R. C., and Brookes, A. J. (2014). Gwas central: a comprehensive resource for the comparison and interrogation of genome-wide association studies. *Eur J Hum Genet*, 22(7):949–52.

Bendl, J., Stourac, J., Salanda, O., Pavelka, A., Wieben, E. D., Zendulka, J., Brezovsky, J., and Damborsky, J. (2014). Predictsnp: robust and accurate consensus classifier for prediction of disease-related mutations. *PLoS Comput Biol*, 10(1):e1003440.

Bernstein, F. C., Koetzle, T. F., Williams, G. J., Meyer, Jr, E. F., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T., and Tasumi, M. (1977). The protein data bank: a computer-based archival file for macromolecular structures. *J Mol Biol*, 112(3):535–42.

Bogan, A. A. and Thorn, K. S. (1998). Anatomy of hot spots in protein interfaces. *J Mol Biol*, 280(1):1–9.

Bromberg, Y., Kahn, P. C., and Rost, B. (2013). Neutral and weakly nonneutral sequence variants may define individuality. *Proc Natl Acad Sci U S A*, 110(35):14255–60.

Bromberg, Y., Overton, J., Vaisse, C., Leibel, R. L., and Rost, B. (2009). In silico mutagenesis: a case study of the melanocortin 4 receptor. *FASEB J*, 23(9):3059–69.

Bromberg, Y. and Rost, B. (2007). Snap: predict effect of non-synonymous polymorphisms on function. *Nucleic Acids Res*, 35(11):3823–35.

Calabrese, R., Capriotti, E., Fariselli, P., Martelli, P. L., and Casadio, R. (2009). Functional annotations improve the predictive score of human disease-related mutations in proteins. *Hum Mutat*, 30(8):1237–44.

Capriotti, E., Fariselli, P., Rossi, I., and Casadio, R. (2008). A three-state prediction of single point mutations on protein stability changes. *BMC Bioinformatics*, 9 Suppl 2:S6.

Clackson, T. and Wells, J. A. (1995). A hot spot of binding energy in a hormone-receptor interface. *Science*, 267(5196):383–6.

Collins, F. S., Brooks, L. D., and Chakravarti, A. (1998). A dna polymorphism discovery resource for research on human genetic variation. *Genome Res*, 8(12):1229–31.

Cook, Jr, E. H. and Scherer, S. W. (2008). Copy-number variations associated with neuropsychiatric conditions. *Nature*, 455(7215):919–23.

Davoli, T. and de Lange, T. (2011). The causes and consequences of polyploidy in normal development and cancer. *Annu Rev Cell Dev Biol*, 27:585–610.

Dehouck, Y., Grosfils, A., Folch, B., Gilis, D., Bogaerts, P., and Rooman, M. (2009). Fast and accurate predictions of protein stability changes upon mutations using statistical potentials and neural networks: Popmusic-2.0. *Bioinformatics*, 25(19):2537–43.

Dehouck, Y., Kwasigroch, J. M., Rooman, M., and Gilis, D. (2013). Beatmusic: Prediction of changes in protein-protein binding affinity on mutations. *Nucleic Acids Res*, 41(Web Server issue):W333–9.

DeLong, E. R., DeLong, D. M., and Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, 44(3):837–45.

Driscoll, D. A. and Gross, S. (2009). Clinical practice. prenatal screening for aneuploidy. *N Engl J Med*, 360(24):2556–62.

Eberle, M. A., Ng, P. C., Kuhn, K., Zhou, L., Peiffer, D. A., Galver, L., Viaud-Martinez, K. A., Lawley, C. T., Gunderson, K. L., Shen, R., and Murray, S. S. (2007). Power to detect risk alleles using genome-wide tag snp panels. *PLoS Genet*, 3(10):1827–37.

ENCODE Project Consortium (2012). An integrated encyclopedia of dna elements in the human genome. *Nature*, 489(7414):57–74.

Fodor, A. A. and Aldrich, R. W. (2004). Influence of conservation on calculations of amino acid covariance in multiple sequence alignments. *Proteins*, 56(2):211–21.

Frank, E., Hall, M., Trigg, L., Holmes, G., and Witten, I. H. (2004). Data mining in bioinformatics using weka. *Bioinformatics*, 20(15):2479–81.

Frazer, K. A., Murray, S. S., Schork, N. J., and Topol, E. J. (2009). Human genetic variation and its contribution to complex traits. *Nat Rev Genet*, 10(4):241–51.

González-Pérez, A. and López-Bigas, N. (2011). Improving the assessment of the outcome of nonsynonymous snvs with a consensus deleteriousness score, condel. *Am J Hum Genet*, 88(4):440–9.

Hamosh, A., Scott, A. F., Amberger, J. S., Bocchini, C. A., and McKusick, V. A. (2005). Online mendelian inheritance in man (omim), a knowledge-base of human genes and genetic disorders. *Nucleic Acids Res*, 33(Database issue):D514–7.

Harewood, L., Schütz, F., Boyle, S., Perry, P., Delorenzi, M., Bickmore, W. A., and Reymond, A. (2010). The effect of translocation-induced nuclear reorganization on gene expression. *Genome Res*, 20(5):554–64.

Hecht, M. (2011). Improve predictions of functional effect of non-synonymous snps. Master's thesis, Technische Universität München.

Hecht, M., Bromberg, Y., and Rost, B. (2013). News from the protein mutability landscape. *J Mol Biol*, 425(21):3937–48.

Hecht, M., Bromberg, Y., and Rost, B. (2015). Better prediction of functional effects for sequence variants. *BMC Genomics*, 16(Suppl 8):S1.

Henikoff, S. and Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A*, 89(22):10915–9.

Hopf, T. A., Colwell, L. J., Sheridan, R., Rost, B., Sander, C., and Marks, D. S. (2012). Three-dimensional structures of membrane proteins from genomic sequencing. *Cell*, 149(7):1607–21.

Iafrate, A. J., Feuk, L., Rivera, M. N., Listewnik, M. L., Donahoe, P. K., Qi, Y., Scherer, S. W., and Lee, C. (2004). Detection of large-scale variation in the human genome. *Nat Genet*, 36(9):949–51.

International HapMap Consortium (2003). The international hapmap project. *Nature*, 426(6968):789–96.

Kabsch, W. and Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22(12):2577–637.

Kawabata, T., Ota, M., and Nishikawa, K. (1999). The protein mutant database. *Nucleic Acids Res*, 27(1):355–7.

Kawashima, S. and Kanehisa, M. (2000). Aaindex: amino acid index database. *Nucleic Acids Res*, 28(1):374.

Kidd, J. M., Cooper, G. M., Donahue, W. F., Hayden, H. S., Sampas, N., Graves, T., Hansen, N., Teague, B., Alkan, C., Antonacci, F., Haugen, E., Zerr, T., Yamada, N. A., Tsang, P., Newman, T. L., Tüzün, E., Cheng, Z., Ebling, H. M., Tusneem, N., David, R., Gillett, W., Phelps, K. A., Weaver, M., Saranga, D., Brand, A., Tao, W., Gustafson, E., McKernan, K., Chen, L., Malig, M., Smith, J. D., Korn, J. M., McCarroll, S. A., Altshuler, D. A., Peiffer, D. A., Dorschner, M., Stamatoyannopoulos, J., Schwartz, D., Nickerson, D. A., Mullikin, J. C., Wilson, R. K., Bruhn, L., Olson, M. V., Kaul, R., Smith, D. R., and Eichler, E. E. (2008). Mapping and sequencing of structural variation from eight human genomes. *Nature*, 453(7191):56–64.

Kiezun, A., Garimella, K., Do, R., Stitziel, N. O., Neale, B. M., McLaren, P. J., Gupta, N., Sklar, P., Sullivan, P. F., Moran, J. L., Hultman, C. M., Lichtenstein, P., Magnusson, P., Lehner, T., Shugart, Y. Y., Price, A. L., de Bakker, P. I. W., Purcell, S. M., and Sunyaev, S. R. (2012). Exome

sequencing and the genetic basis of complex traits. *Nat Genet*, 44(6):623–30.

Kimura, M. (1968). Evolutionary rate at the molecular level. *Nature*, 217(5129):624–6.

Kircher, M., Witten, D. M., Jain, P., O'Roak, B. J., Cooper, G. M., and Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet*, 46(3):310–5.

Kowarsch, A., Fuchs, A., Frishman, D., and Pagel, P. (2010). Correlated mutations: a hallmark of phenotypic amino acid substitutions. *PLoS Comput Biol*, 6(9).

Krishnamurthy, K. (2003). *Textbook of Biodiversity*. Taylor & Francis.

Kruglyak, L. and Nickerson, D. A. (2001). Variation is the spice of life. *Nat Genet*, 27(3):234–6.

Kumar, P., Henikoff, S., and Ng, P. C. (2009). Predicting the effects of coding non-synonymous variants on protein function using the sift algorithm. *Nat Protoc*, 4(7):1073–81.

Lander, E. S. (2011). Initial impact of the sequencing of the human genome. *Nature*, 470(7333):187–97.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J. P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P.,

Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J. C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R. H., Wilson, R. K., Hillier, L. W., McPherson, J. D., Marra, M. A., Mardis, E. R., Fulton, L. A., Chinwalla, A. T., Pepin, K. H., Gish, W. R., Chissoe, S. L., Wendl, M. C., Delehaunty, K. D., Miner, T. L., Delehaunty, A., Kramer, J. B., Cook, L. L., Fulton, R. S., Johnson, D. L., Minx, P. J., Clifton, S. W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J. F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., Gibbs, R. A., Muzny, D. M., Scherer, S. E., Bouck, J. B., Sodergren, E. J., Worley, K. C., Rives, C. M., Gorrell, J. H., Metzker, M. L., Naylor, S. L., Kucherlapati, R. S., Nelson, D. L., Weinstock, G. M., Sakaki, Y., Fujiyama, A., Hattori, M., Yada, T., Toyoda, A., Itoh, T., Kawagoe, C., Watanabe, H., Totoki, Y., Taylor, T., Weissenbach, J., Heilig, R., Saurin, W., Artiguenave, F., Brottier, P., Bruls, T., Pelletier, E., Robert, C., Wincker, P., Smith, D. R., Doucette-Stamm, L., Rubenfield, M., Weinstock, K., Lee, H. M., Dubois, J., Rosenthal, A., Platzer, M., Nyakatura, G., Taudien, S., Rump, A., Yang, H., Yu, J., Wang, J., Huang, G., Gu, J., Hood, L., Rowen, L., Madan, A., Qin, S., Davis, R. W., Federspiel, N. A., Abola, A. P., Proctor, M. J., Myers, R. M., Schmutz, J., Dickson, M., Grimwood, J., Cox, D. R., Olson, M. V., Kaul, R., Raymond, C., Shimizu, N., Kawasaki, K., Minoshima, S., Evans, G. A., Athanasiou, M., Schultz, R., Roe, B. A., Chen, F., Pan, H., Ramser, J., Lehrach, H., Reinhardt, R., McCombie, W. R., de la Bastide, M., Dedhia, N., Blöcker, H., Hornischer, K., Nordsiek, G., Agarwala, R., Aravind, L., Bailey, J. A., Bateman, A., Batzoglou, S., Birney, E., Bork, P., Brown, D. G., Burge, C. B., Cerutti, L., Chen, H. C., Church, D., Clamp, M., Copley, R. R., Doerks, T., Eddy, S. R., Eichler, E. E., Furey, T. S., Galagan, J., Gilbert, J. G., Harmon, C., Hayashizaki, Y., Haussler,

D., Hermjakob, H., Hokamp, K., Jang, W., Johnson, L. S., Jones, T. A., Kasif, S., Kaspryzk, A., Kennedy, S., Kent, W. J., Kitts, P., Koonin, E. V., Korf, I., Kulp, D., Lancet, D., Lowe, T. M., McLysaght, A., Mikkelsen, T., Moran, J. V., Mulder, N., Pollara, V. J., Ponting, C. P., Schuler, G., Schultz, J., Slater, G., Smit, A. F., Stupka, E., Szustakowski, J., Thierry-Mieg, D., Thierry-Mieg, J., Wagner, L., Wallis, J., Wheeler, R., Williams, A., Wolf, Y. I., Wolfe, K. H., Yang, S. P., Yeh, R. F., Collins, F., Guyer, M. S., Peterson, J., Felsenfeld, A., Wetterstrand, K. A., Patrinos, A., Morgan, M. J., de Jong, P., Catanese, J. J., Osoegawa, K., Shizuya, H., Choi, S., Chen, Y. J., Szustakowki, J., and International Human Genome Sequencing Consortium (2001). Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921.

Liu, X., Jian, X., and Boerwinkle, E. (2011). dbnsfp: a lightweight database of human nonsynonymous snps and their functional predictions. *Hum Mutat*, 32(8):894–9.

Loeb, D. D., Swanstrom, R., Everitt, L., Manchester, M., Stamper, S. E., and Hutchison, 3rd, C. A. (1989). Complete mutagenesis of the hiv-1 protease. *Nature*, 340(6232):397–400.

Markiewicz, P., Kleina, L. G., Cruz, C., Ehret, S., and Miller, J. H. (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(5):421–33.

Marks, D. S., Hopf, T. A., and Sander, C. (2012). Protein structure prediction from sequence variation. *Nat Biotechnol*, 30(11):1072–80.

Martinez-Perez, E., Shaw, P., Aragon-Alcaide, L., and Moore, G. (2003). Chromosomes form into seven groups in hexaploid and tetraploid wheat as a prelude to meiosis. *Plant J*, 36(1):21–9.

McCarthy, M. I., Abecasis, G. R., Cardon, L. R., Goldstein, D. B., Little, J., Ioannidis, J. P. A., and Hirschhorn, J. N. (2008). Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat Rev Genet*, 9(5):356–69.

Ng, P. C., Henikoff, J. G., and Henikoff, S. (2000). Phat: a transmembrane-specific substitution matrix. predicted hydrophobic and transmembrane. *Bioinformatics*, 16(9):760–6.

Nissen, S. (2003). Implementation of a fast artificial neural network library (fann). *Report, Department of Computer Science University of Copenhagen (DIKU)*, 31.

Ofran, Y., Mysore, V., and Rost, B. (2007). Prediction of dna-binding residues from sequence. *Bioinformatics*, 23(13):i347–53.

Ofran, Y. and Rost, B. (2007). Isis: interaction sites identified from sequence. *Bioinformatics*, 23(2):e13–6.

Ogura, Y., Bonen, D. K., Inohara, N., Nicolae, D. L., Chen, F. F., Ramos, R., Britton, H., Moran, T., Karaliuskas, R., Duerr, R. H., Achkar, J. P., Brant, S. R., Bayless, T. M., Kirschner, B. S., Hanauer, S. B., Nuñez, G., and Cho, J. H. (2001). A frameshift mutation in nod2 associated with susceptibility to crohn's disease. *Nature*, 411(6837):603–6.

Ohta, T. (2002). Near-neutrality in evolution of genes and gene regulation. *Proc Natl Acad Sci U S A*, 99(25):16134–7.

Patterson, D. (2009). Molecular genetic analysis of down syndrome. *Hum Genet*, 126(1):195–214.

Pe'er, I., de Bakker, P. I. W., Maller, J., Yelensky, R., Altshuler, D., and Daly, M. J. (2006). Evaluating and improving power in whole-genome association studies using fixed marker sets. *Nat Genet*, 38(6):663–7.

Perry, G. H., Dominy, N. J., Claw, K. G., Lee, A. S., Fiegler, H., Redon, R., Werner, J., Villanea, F. A., Mountain, J. L., Misra, R., Carter, N. P., Lee, C., and Stone, A. C. (2007). Diet and the evolution of human amylase gene copy number variation. *Nat Genet*, 39(10):1256–60.

Punta, M., Coggill, P. C., Eberhardt, R. Y., Mistry, J., Tate, J., Boursnell, C., Pang, N., Forslund, K., Ceric, G., Clements, J., Heger, A., Holm, L., Sonnhammer, E. L. L., Eddy, S. R., Bateman, A., and Finn, R. D. (2012). The pfam protein families database. *Nucleic Acids Res*, 40(Database issue):D290–301.

Rauch, A., Wieczorek, D., Graf, E., Wieland, T., Endele, S., Schwarzmayr, T., Albrecht, B., Bartholdi, D., Beygo, J., Di Donato, N., Dufke, A., Cremer, K., Hempel, M., Horn, D., Hoyer, J., Joset, P., Röpke, A., Moog, U., Riess, A., Thiel, C. T., Tzschach, A., Wiesener, A., Wohlleber, E., Zweier, C., Ekici, A. B., Zink, A. M., Rump, A., Meisinger, C., Grallert, H., Sticht, H., Schenck, A., Engels, H., Rappold, G., Schröck, E., Wieacker, P., Riess, O., Meitinger, T., Reis, A., and Strom, T. M. (2012). Range of genetic mutations associated with severe non-syndromic sporadic intellectual disability: an exome sequencing study. *Lancet*, 380(9854):1674–82.

Reva, B., Antipin, Y., and Sander, C. (2011). Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic Acids Res*, 39(17):e118.

Rosenbaum, D. M., Zhang, C., Lyons, J. A., Holl, R., Aragao, D., Arlow, D. H., Rasmussen, S. G. F., Choi, H.-J., Devree, B. T., Sunahara, R. K., Chae, P. S., Gellman, S. H., Dror, R. O., Shaw, D. E., Weis, W. I., Caffrey, M., Gmeiner, P., and Kobilka, B. K. (2011). Structure and function of an irreversible agonist-beta(2) adrenoceptor complex. *Nature*, 469(7329):236–40.

Rost, B. (1999). Twilight zone of protein sequence alignments. *Protein Eng*, 12(2):85–94.

Rost, B. and Sander, C. (1993). Prediction of protein secondary structure at better than 70% accuracy. *J Mol Biol*, 232(2):584–99.

Rost, B. and Sander, C. (1994). Conservation and prediction of solvent accessibility in protein families. *Proteins*, 20(3):216–26.

Sachidanandam, R., Weissman, D., Schmidt, S. C., Kakol, J. M., Stein, L. D., Marth, G., Sherry, S., Mullikin, J. C., Mortimore, B. J., Willey, D. L., Hunt, S. E., Cole, C. G., Coggill, P. C., Rice, C. M., Ning, Z., Rogers, J., Bentley, D. R., Kwok, P. Y., Mardis, E. R., Yeh, R. T., Schultz, B., Cook, L., Davenport, R., Dante, M., Fulton, L., Hillier, L., Waterston, R. H., McPherson, J. D., Gilman, B., Schaffner, S., Van Etten, W. J., Reich, D., Higgins, J., Daly, M. J., Blumenstiel, B., Baldwin, J., Stange-Thomann, N., Zody, M. C., Linton, L., Lander, E. S., Altshuler, D., and International SNP Map Working Group (2001). A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature*, 409(6822):928–33.

Sander, C. and Schneider, R. (1991). Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins*, 9(1):56–68.

Sankararaman, S., Sha, F., Kirsch, J. F., Jordan, M. I., and Sjölander, K. (2010). Active site prediction using evolutionary and structural information. *Bioinformatics*, 26(5):617–24.

Sankararaman, S. and Sjölander, K. (2008). Intrepid–information-theoretic tree traversal for protein functional site identification. *Bioinformatics*, 24(21):2445–52.

Schaefer, C., Bromberg, Y., Achten, D., and Rost, B. (2012). Disease-related mutations predicted to impact protein function. *BMC Genomics*, 13 Suppl 4:S11.

Schaefer, C. and Rost, B. (2012). Predict impact of single amino acid change upon protein structure. *BMC Genomics*, 13 Suppl 4:S4.

Schlessinger, A., Yachdav, G., and Rost, B. (2006). Profbval: predict flexible and rigid residues in proteins. *Bioinformatics*, 22(7):891–3.

Schwarz, J. M., Rödelsperger, C., Schuelke, M., and Seelow, D. (2010). Mutationtaster evaluates disease-causing potential of sequence alterations. *Nat Methods*, 7(8):575–6.

Sebat, J., Lakshmi, B., Troge, J., Alexander, J., Young, J., Lundin, P., Månér, S., Massa, H., Walker, M., Chi, M., Navin, N., Lucito, R., Healy, J., Hicks, J., Ye, K., Reiner, A., Gilliam, T. C., Trask, B., Patterson, N., Zetterberg, A., and Wigler, M. (2004). Large-scale copy number polymorphism in the human genome. *Science*, 305(5683):525–8.

Sen, S. (2000). Aneuploidy and cancer. *Curr Opin Oncol*, 12(1):82–8.

Sherry, S. T., Ward, M. H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., and Sirotkin, K. (2001). dbsnp: the ncbi database of genetic variation. *Nucleic Acids Res*, 29(1):308–11.

Sigrist, C. J. A., Cerutti, L., de Castro, E., Langendijk-Genevaux, P. S., Bulliard, V., Bairoch, A., and Hulo, N. (2010). Prosite, a protein domain database for functional characterization and annotation. *Nucleic Acids Res*, 38(Database issue):D161–6.

Slatkin, M. (2008). Linkage disequilibrium–understanding the evolutionary past and mapping the medical future. *Nat Rev Genet*, 9(6):477–85.

Stankiewicz, P. and Lupski, J. R. (2010). Structural variation in the human genome and its role in disease. *Annu Rev Med*, 61:437–55.

Stenson, P. D., Ball, E. V., Mort, M., Phillips, A. D., Shiel, J. A., Thomas, N. S. T., Abeysinghe, S., Krawczak, M., and Cooper, D. N. (2003). Human gene mutation database (hgmd): 2003 update. *Hum Mutat*, 21(6):577–81.

Sunyaev, S. R., Eisenhaber, F., Rodchenkov, I. V., Eisenhaber, B., Tumanyan, V. G., and Kuznetsov, E. N. (1999). Psic: profile extraction from sequence alignments with position-specific counts of independent observations. *Protein Eng*, 12(5):387–94.

Teer, J. K. and Mullikin, J. C. (2010). Exome sequencing: the sweet spot before whole genomes. *Hum Mol Genet*, 19(R2):R145–51.

Thorn, K. S. and Bogan, A. A. (2001). Asedb: a database of alanine mutations and their effects on the free energy of binding in protein interactions. *Bioinformatics*, 17(3):284–5.

Thusberg, J. and Vihinen, M. (2009). Pathogenic or not? and if so, then how? studying the effects of missense mutations using bioinformatics methods. *Hum Mutat*, 30(5):703–14.

UniProt Consortium (2015). Uniprot: a hub for protein information. *Nucleic Acids Res*, 43(Database issue):D204–12.

Webb, E. C. (1992). *Enzyme nomenclature 1992. Recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the Nomenclature and Classification of Enzymes.* Number Ed. 6. Academic Press, New York.

Wellcome Trust Case Control Consortium (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145):661–78.

Yachdav, G., Hecht, M., Pasmanik-Chor, M., Yeheskel, A., and Rost, B. (2014). Heatmapviewer: interactive display of 2d data in biology. *F1000Res*, 3:48.

Zhang, F., Gu, W., Hurles, M. E., and Lupski, J. R. (2009). Copy number variation in human health, disease, and evolution. *Annu Rev Genomics Hum Genet*, 10:451–81.

Appendix

This work constitutes a cumulative dissertation and the methodologies and results presented in this thesis have been published in peer-reviewed journals. The corresponding articles are appended to this dissertation and will be briefly summarized in the following:

- **Maximilian Hecht**, Yana Bromberg, and Burkhard Rost (2013). **News from the protein mutability landscape**. Journal of Molecular Biology, 425(21):3937-48.
- **Maximilian Hecht**, Yana Bromberg, and Burkhard Rost (2015). **Better prediction of functional effects for sequence variants**. BMC Genomics, 16(Suppl 8):S1.

News from the protein mutability landscape

In this publication we present a novel approach towards variant effect prediction by expanding the view from single variants to complete *in-silico* mutagenesis predictions. We widely review the field of variant effect prediction by presenting methods and discussing the current state-of-the-art. We show that methods are able to computationally predict the effects of most variants and that effect predictions tend to cluster around important functional regions. Although most methods perform comparably there is a significant difference between individual methods' predictions, which are attributable to the features and data used in method development. In this article we moreover present functional effect predictions for a complete *in-silico* mutagenesis experiment in the human beta-2-adrenergic receptor. We show the generated predictions as a heat map and how these correlate with secondary structure elements in the protein. To further visualize our results, we used the known experimental 3D structure and highlighted it according to our predictions thus making high-impact regions in the protein immediately visible in the structure. This example shows how computational predictions of the mutability landscape can translate into novel hypotheses on protein function and possibly aid drug development.

Author contributions: Maximilian Hecht, Yana Bromberg and Burkhard Rost conceived the study design and methodologies. Maximilian Hecht collected the necessary data and carried out the experiments. Results were analyzed by Maximilian Hecht, Yana Bromberg and Burkhard Rost. The manuscript was written, revised and approved by Maximilian Hecht, Yana Bromberg and Burkhard Rost.

Better prediction of functional effects for sequence variants

This article describes the development and evaluation of our novel method SNAP2, a machine learning-based classifier for the prediction of functional effects of sequence variants. We show that SNAP2 compares favorably with other state-of-the-art methods and that it is capable of producing non-random predictions for orphan proteins. Our analysis shows that certain variants cannot be predicted equally well by different methods and that our novel method performs remarkably well on variants where multiple methods disagree. Moreover, we show that a naïve method combination, that is often employed by biologists and geneticists, performs worse than the individual methods on our data. We present the improvements made to our method and discuss how these affect the results in different testing scenarios. We also discuss how data bias affects computational predictions of variant effects and what users can expect from using these methods.

Author contributions: Maximilian Hecht, Yana Bromberg and Burkhard Rost conceived this work and designed the experiments. Maximilian Hecht wrote the software and carried out the experiments. Maximilian Hecht and Yana Bromberg collected the data and analyzed the results together with Burkhard Rost. The manuscript was written, revised and approved by Maximilian Hecht, Yana Bromberg and Burkhard Rost.

News from the Protein Mutability Landscape

Maximilian Hecht[1], **Yana Bromberg**[2] **and Burkhard Rost**[1,3,4]

*1 - **Department of Bioinformatics and Computational Biology I12,** Technische Universität München, Boltzmannstrasse 3, 85748 Garching, Germany*
*2 - **Department of Biochemistry and Microbiology,** Rutgers University, 76 Lipman Drive, New Brunswick, NJ 08901, USA*
*3 - **Institute of Advanced Study,** Technische Universität München, Boltzmannstrasse 3, 85748 Garching, Germany*
*4 - **Institute for Food and Plant Sciences,** Life Science Center Weihenstephan, Alte Akademie 8, 85354 Freising, Germany*

Correspondence to Maximilian Hecht: *hecht@rostlab.org*
http://dx.doi.org/10.1016/j.jmb.2013.07.028
Edited by E. Alexov

Abstract

Some mutations of protein residues matter more than others, and these are often conserved evolutionarily. The explosion of deep sequencing and genotyping increasingly requires the distinction between effect and neutral variants. The simplest approach predicts all mutations of conserved residues to have an effect; however, this works poorly, at best. Many computational tools that are optimized to predict the impact of point mutations provide more detail. Here, we expand the perspective from the view of single variants to the level of sketching the entire mutability landscape. This landscape is defined by the impact of substituting every residue at each position in a protein by each of the 19 non-native amino acids. We review some of the powerful conclusions about protein function, stability and their robustness to mutation that can be drawn from such an analysis. Large-scale experimental and computational mutagenesis experiments are increasingly furthering our understanding of protein function and of the genotype–phenotype associations. We also discuss how these can be used to improve predictions of protein function and pathogenicity of missense variants.

Introduction

> *Lewis Carroll—Through the Looking Glass: "Why, sometimes I've believed as many as six impossible things before breakfast."*

Understanding Genetic Diversity—A Central Challenge for Deep Sequencing

Elucidating which human genetic variations have which phenotypic effect and how the variation impacts disease is one of the major scientific challenges in the 21st century. While the vast majority of genetic variants are hypothesized to be neutral [1], that is, are assumed not to contribute to any phenotype, the relative percentage of neutral, near-neutral [2] and non-neutral variants remains unclear. Most likely, the precise ratios heavily depend on the particular protein under investigation (e.g., the human immunodeficiency virus gp120 is likely to be much more robust against mutation than p53 simply because many of the p53 residues are involved in binding and therefore "vulnerable" to mutation). A key aspect in the development of strategies for diagnosis and treatment of genetic diseases is to further our understanding of the underlying mechanisms that link genotypes and phenotypes.

Sequence variants such as single nucleotide polymorphisms (SNPs) are the most prevalent form of human genetic variation [3]. It has been estimated that more than 11 million SNPs will be observed among people; 7 million of these are frequent (common variants), that is, occur with a minor allele frequency above 5%, while the remaining (minor allele frequency, $<5\%$) are considered as rare [4]. Many of both rare and common variants may be instrumental in defining individual's differences [5–7]. Increasingly, however,

researchers begin to suspect that every possible point mutation might ultimately be observed.

For medical biology, non-synonymous SNPs (nsSNPs) or missense variants that change the amino acid sequence of the protein are particularly interesting. These variants are more likely to affect function than synonymous SNPs. Single-amino-acid variants can change the resulting phenotype, for example, by altering protein function directly or indirectly by impacting structure and/or binding. Such changes can lead to pathogenic phenotypes [8]. Recent studies suggest that every pair of individuals differs by almost one amino acid variant in each protein while individuals have about 1.2–1.7 variants (nsSNPs) that are novel with respect to both parents, that is, not observed in either parent [7].

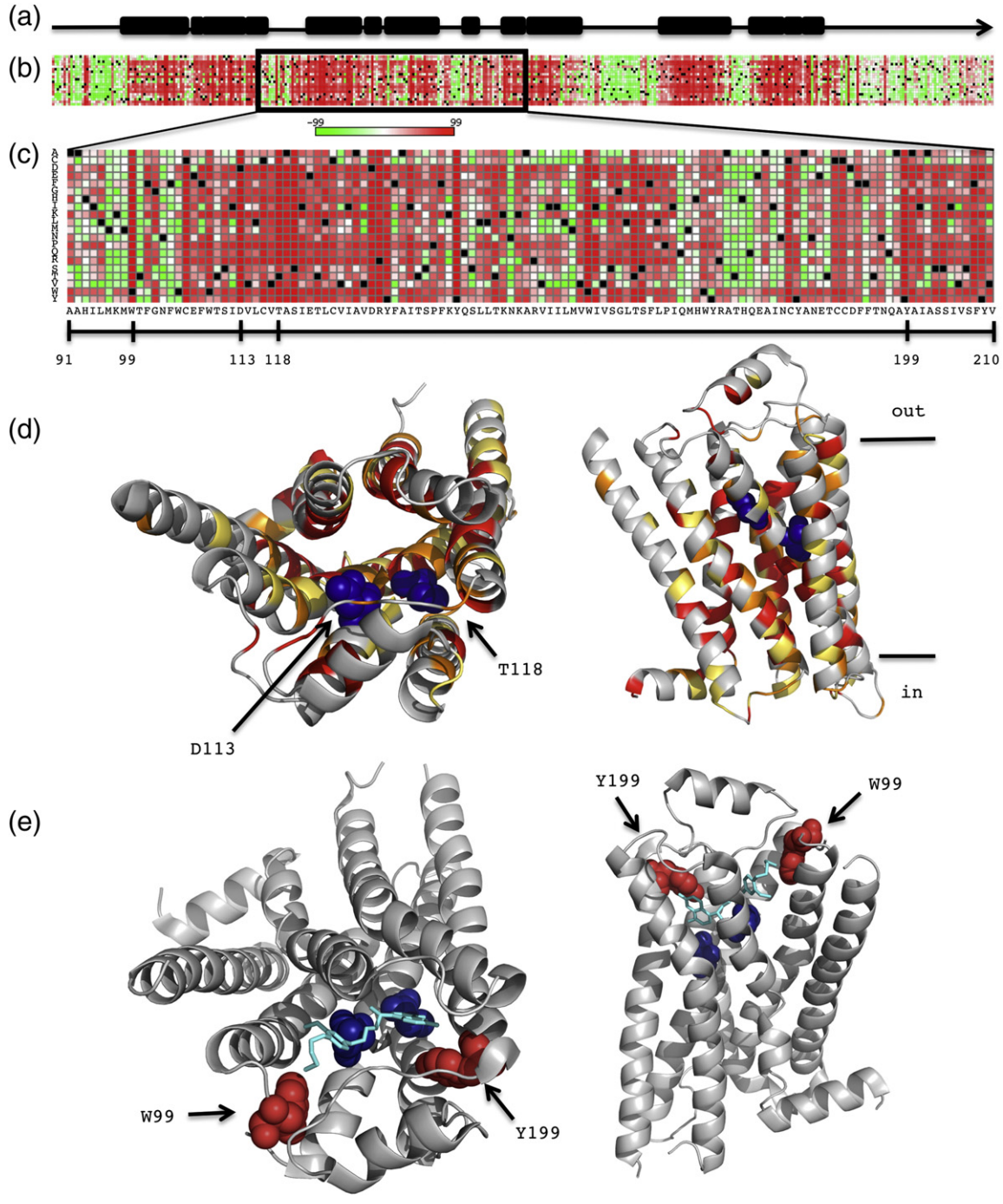

Fig. 1 (*legend on next page*)

Knowing how these changes affect function can give, for instance, insight into a child's disease predisposition.

GWAS (*g*enome-*w*ide *a*ssociation *s*tudies) has evolved as the most widely used approach relating human genetic variation to phenotypic diversity [9]. Results of these studies greatly increased our understanding of molecular pathways underlying specific human diseases. Most common SNPs have been assessed for statistical associations with many complex traits and common diseases. However, for the vast majority of complex trait associations, the underlying mechanisms remain unknown, and for many of the known common SNPs related to complex disease phenotypes, GWAS misses the known associations [10]. Rare variants are missed due to limited numbers [4,11]. The hope is that deep sequencing will address some of these issues by revealing variations between the full sequences.

Mutability Landscapes May Be THE Key to Understanding Diversity

Meanwhile, a different approach toward understanding the genotype–phenotype association is to study functionally important regions, robustness and evolvability of proteins by investigating the mutability landscape. The mutability landscape of, for example, protein function, can be defined as the effect of all possible point mutations/variants upon protein function, that is, of substituting the native amino acid at each residue position against all 19 non-native amino acids, one at a time (Fig. 1 [20,21]). Studying such a landscape may help us a lot in understanding protein function and evolution.

In this review, we exclusively focus on the effects of varying the protein sequence by single-amino-acid substitutions (SAASs). In order to avoid obscuring acronyms, we will simply use the term variant as synonym for SAAS. We review comprehensive mutagenesis in which each position in a protein is changed and complete mutagenesis in which each position is replaced by every non-native amino acid. However, we largely discard effects of varying multiple residues at the same time. Obviously, even such a reduced version of the *protein mutability* landscape already carries very important information about protein function. We attempt to sketch how this landscape brings about new challenges and new possibilities. In fact, we have to learn to understand what we see in this new looking glass.

Most Variant Effects Predicted Correctly *In Silico*

Several computational methods predict the effect of variants (SAAS or nsSNPs). Some predict the effect on protein function {e.g., sorting intolerant from tolerant (SIFT) [22,23] or screening for non-acceptable polymorphisms (SNAP) [24,25]}, others predict the effect with respect to their pathogenicity (e.g., MutPred [26], SNPs&GO [27], Mutation Assesor [28] or MutationTaster [29]) and others yet predict the effect on protein structure directly [30] or cannot be easily fit into these categories (e.g., PolyPhen-2 [31] or PON-P [32]). These methods use a diverse spectrum of input features, typically combining evolutionary information with biophysical features and experimental information about protein structure and function where available. There are several outstanding reviews on the prediction of functional effects [33–35], and the community puts great effort into assessing such predictions. For instance, CAGI (*C*ritical *A*ssessment of *G*enome *I*nterpretation) aspires to assessing method performance in predicting phenotypic impacts of genomic variation [36]. More formal studies assess predictors specifically with respect to their performance in identifying pathogenic variants [37]. The results of these studies suggest that each method has strengths and weaknesses, possibly resulting from the data used for development and the types of information

Fig. 1. Mutability landscape of a protein. The top line (a) sketches the sequence and secondary structure (transmembrane helices) of the adrenergic receptor (ADRB2_HUMAN, ID: P07550 [12]; assignment of secondary from the high-resolution structure PDB ID: 3PDS [13] using DSSP [14]). For each of the 413 residues (*x*-axis) of the receptor, (b) shows the predictions for the effects of all 19 non-native variants (*y*-axis; the stronger the predicted effect, the redder; the stronger the predicted neutrality, the greener). (c) Zoom of the fragment spanning from residue 91 to residue 210 and the relative positions of binding sites (D113 and T118) and proposed target residues (W99 and Y199). (d) The predicted functional effect of variants for the 3D structure (PDB ID: 2RH1 [15]); both known binding sites (positions 113 and 118) are shown as blue spheres. Shown are the average scores [SNAP score ranges from the most neutral (−100) to the strongest effect (+100)] over amino acids that would be considered as "neutral" given the biophysical amino acid features as captured in the PHAT substitution matrix [16] for transmembrane regions and in the BLOSUM62 matrix [17] for all other residues. Red depicts high average scores (score > 60), orange depicts intermediate scores (40 < score < 60), yellow depicts low scores (20 < score < 40) and gray marks sites with SNAP scores <20 (predicted as neutral or with little effect). (e) The 3D structure (PDB ID: 3PDS [13]) with a bound agonist and the two residues (W99 and Y199) that exhibit a high overall predicted effect and are under strong evolutionary constraint (predicted by EVfold [18,19]) with each other and the two binding sites.

included in prediction. Good *in silico* methods correctly predict the experimentally observed effects for most variants.

Typically, only 5–10% of all residues relate directly to function [38–40]. Some of these are revealed in the substitution profiles of protein families. A whole generation of methods targets the prediction of such functional sites through analyzing evolutionary information (e.g., ET [41], INTREPID [42] or DISCERN [43]). Since these methods predict functional sites, it is not surprising that they also capture some of the signal that variants impact function (V. Link and K. Sjölander, unpublished results). Thus, the ability to predict the functional effect of variants is clearly related to predicting protein function.

Predictions of variant effects have helped us prioritize mutations for large-scale reverse genetics projects, where mutations are randomly introduced into the genome. An example for such strategies is TILLING (*t*argeting *i*nduced *l*ocal *l*esions *in g*enomes), a method that combines chemical mutagenesis with a sensitive DNA screening technique in order to allow direct identification of mutations in a specific gene. TILLING uses the functional effect predictions from SIFT [22,23] to prioritize the post-processing of variants [44]. Another important application is to the assessment of disease-related human variants [20,45,46]. For instance, mutations that directly cause a disease, such as those found in OMIM [47], are clearly identified by methods that predict the functional effect of variants [48]. Existing *in silico* methods can even be good enough to reveal problems with experimental data: today's assessment of functional neutrality of variants seems particularly problematic [100]

Peeking Experimentally into the Protein Mutability Landscape

Alanine scans reveal function and interaction hot spots

The experimental study on how site-directed mutagenesis affects phenotypes may be THE most essential experimental tool for determining protein function. By substituting residues that are assumed to be important and measuring substitution effect, researchers identify the residues that are important for the hypothesized protein function. Over the last decade, the power of experimental and computational mutagenesis has grown considerably: a decade ago, many publications reported on single point mutations; today, 50 times more may no longer satisfy reviewers.

The ability of proteins to interact with substrates or other proteins is essential. The most important function of a protein can therefore be defined as its role within an interaction pathway [49]. Typically, only a few residues in a protein interaction interface contribute most of the binding affinity. These can be identified by the change in binding free energy upon mutation to alanine and are often referred to as binding *hot spots* [50,51]. One definition for a hot spot is that the binding free energy is altered by $\geq$1 kcal/mol upon mutation [52]. While the precise definition might be subject to debate, hot spots are "real" in the sense that they can be predicted accurately by methods that do not even assume that hot spots exist [53]. We moreover might anticipate the observation that the residues or positions contributing most to the energy of binding might also be the residues used more frequently when choosing sites that bind to many binding partners. Indeed, hot spots have been observed to have a high propensity for interaction with multiple partners [54].

Substituting the native amino acid by alanine is typically experimentally easiest and expected to be most revealing. Thus, alanine scans are most common, but increasingly, glycine, proline and cysteine scans are also carried out [55–57]. In these scans, all native residues of a protein are individually substituted by one of the above amino acids and the effect upon a given functional assay is measured. ASEdb, the Alanine Scanning Energetics database, provides a central repository for such data [58]. Residues that significantly change protein function are usually considered important. What constitutes a significant change depends on the type of function. *In silico* predictions suggest that when looking at the effect of all 19-non-native mutations, alanine substitutions are most representative (correlate most with the average over all mutations) [21]. Although this observation is based on one single protein (HXK4) and may therefore not be representative, the fact that *in silico* methods accurately reproduce such expert knowledge should be appreciated as an independent evidence of their success in predicting essential aspects of the mutability landscape.

Mutability landscape constrained by correlation networks?

Comprehensive experimental mutagenesis studies confirmed that the effect of point mutations (SAAS) upon function depends crucially on their positions in the protein sequence [59–61]. Even within a unit as familiar as the DNA binding domain of the *Escherichia coli* LacI [62,63] repressor, almost any variant can be tolerated at some positions while, at others, all variants affect function (Fig. 2a). Simple structural constraints might suffice to explain this variability: to accommodate the negatively charged DNA, binding regions of the repressor contain positively charged residues. Furthermore, binding

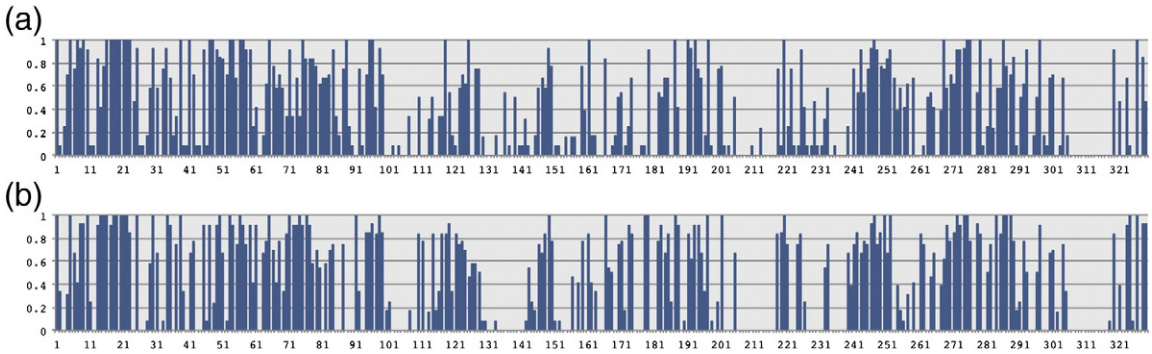

Fig. 2. Mutagenesis of *E. coli* LacI repressor. At each position between residue 2 and 329, 12–13 amino acid substitutions are displayed as a bar. The height of a bar depicts the relative percentage of substitutions that alter the repressor function as determined (a) experimentally [59] or (b) by computational prediction using SNAP2. With a correlation of 0.76 over all residues and an accuracy of 78.2% over all variants, this constitutes a below-average prediction of SNAP2 (~82% estimated overall accuracy).

requires helix formation. These two simple biophysical realities constrain the mutability landscape significantly in a specific, identity-revealing manner like a fingerprint. The differential sensitivity to mutation might just be a complex overlay of many such simple biophysical constraints.

The same constraints are written into the profile of evolutionary conservation of changes observed within families of related proteins [64–66]. These evolutionary imprints are strong enough to aid the prediction of protein structure [67–70] and function [39]. One particular idea that uses the constraints imposed by the mutability landscape is that of compensating/correlated mutations [71–73]. To simplify this, imagine a salt bridge, that is, the interaction between a positively charged residue and a negatively charged residue. If the negative one is mutated into a positively charged amino acid, the affected protein may malfunction. A compensatory mutation that also flips the charge of the positive position will again allow salt-bridge formation. If we could identify correlated mutations, we could use them to predict inter-residue contacts within [72] and between proteins [74,75]. After many years of development [18,76], this concept has finally brought about *de novo* predictions for three-dimensional (3D) structure of globular [77] and large membrane proteins [18,19], several of those are not similar to any protein structure we know today. Such predictions may even lead the way beyond structure [19]: many residues that are evolutionarily coupled and not close in space may be relevant for protein function. In fact, in a study of over 14,000 variants related to disease from over 1000 human proteins, correlated positions appeared significantly more likely to harbor disease mutations than average positions [78]. Compensatory mutations involve coupled variants and might rashly be considered to go beyond the focus of this review. We show that the correlated mutations perfectly highlight the importance of analyzing the mutability landscape.

Other recent studies carry the theme of coevolving positions even further. Patterns of correlated mutations in the WW domain nearly suffice to synthetize artificial WW domains with native-like folding and function [79,80]. Applying statistical coupling analysis to the S1A protein family, the Ranganathan laboratory introduced the "sector hypothesis" [81] that proteins are organized into distinct subunits or networks (sectors) of coevolving residues that are essential to structure and function. Such sectors involve only about every fifth residue; they are built around active sites, and they connect to other functional sites distant in sequence and structure through "networks" of contiguous residue interactions in the protein core [82]. These networks of coevolving residues may have resulted from the need for rapid adaptive variation arising from fluctuating selection pressure and that the organization into networks of cooperatively acting residues may provide such rapid adaptive potential through only a few mutations [82]. If so, structure and function may mostly be affected by mutations at sector positions while non-sector positions may tolerate variation. This hypothesis was tested through a complete single mutagenesis (individually substituting each residue by all 19 non-native amino acids) in one representative member of the PDZ family ($PSD95^{pdz3}$) [82]. The study showed that the statistical correlation between mutations with significant functional effect and sector positions was very strong; it was, in fact, stronger than that between mutations in the protein core (buried positions) and positions with ligand contacts. Moreover, a combination of two mutations at sector positions was sufficient to change the binding specificity of $PSD95^{pdz3}$ for a class-switching ligand. This adaption is exclusively initiated through mutations in the sector. While awaiting large-scale confirmation,

these findings already highlight the importance of annotating correlated mutational behavior for the prediction of pathogenicity/functional effects of missense variants.

Penetrating the Protein Mutability Landscape *In Silico*

Predicting the mutability landscape of the human exome

Ultimately, we want to study the entire protein mutability landscapes for at least some hundreds of representative proteins by assaying changes in protein function and their impact upon the organism. Despite tremendous breakthroughs in high-throughput experimentation, this analysis falls more into the world of Lewis Carroll than into that of a scientific grant proposal. However, such a landscape can be easily predicted, for example, for all human proteins. The downside is that we do not yet fully understand how to interpret the results. Nevertheless, in the context of understanding the deep sequencing data, such views are needed.

Finding the causal variants for a particular disease continues to be a challenging endeavor despite the continued decrease of the cost in sequencing entire genomes and entire exomes [83]. Accordingly, researches prioritize zooming in onto candidate variants in these studies by including computational effect predictions [84]. It has been suggested to combine several prediction methods (e.g., through majority vote) in order to overcome individual weaknesses and obtain most reliable predictions [85]. The dbNSFP [86] database is built to simplify this endeavor by providing effect predictions and scores from various methods for every potential variant in the human genome (approximately 76 million variants). Differences between methods become most apparent when comparing predictions on this large scale. The pairwise agreement between the four methods in dbNSFP ranges from 61% to 77%. The fraction of all potential substitutions predicted to be deleterious by individual methods ranges from 40% to 56%, suggesting that methods disagree strongly. Overall, methods accurately predict Mendelian disease-causing variants to strongly effect function. Unfortunately, this does not imply that the same methods can find a single disease-causing variant among the thousands of variants observed between any pair of individuals from the same population.

To visualize the predictive behavior of the two widely used methods SNAP [24,25] and SIFT [22,23], we compiled a pairwise amino acid substitution matrix over all theoretically possible variants in the human proteome based on the predictions of each method (Fig. 3; note that not all of those variants can be observed since not all amino acids can be transformed into all others through a SNP). Although SNAP predicts more effect substitutions than SIFT, trends appear to be largely similar. For instance, both methods predict substitutions from and to tryptophan as highly damaging on average. This is plausible due to its structural importance to proteins. Similarly, both methods predict substitutions of phenylalanine by any other residue, except for leucine and tyrosine, as rather damaging. Phenylalanine is preferentially exchanged with tyrosine, which differs only in that it contains a hydroxyl group in place of the *ortho* hydrogen on the benzene ring. The preference for leucine seems plausible due to its hydrophobic character. Examples of SNAP and SIFT differences are in the predictions for substitutions of arginine and by proline. SNAP might be closer to the truth for these two because they may be difficult to treat via a purely evolution based method. “To proline” mutations are likely to be rare due to their disruptions. For arginine, the explanation seems less clear. How can we cast such predictions into new methods that, for example, predict active sites? How can we use them to guide protein design?

Outcome of alanine scans predicted

Methods that predict functional effects have rarely been assessed in large-scale mutagenesis experiments. One reason is obviously the shortage of such experiments. Another might be the perception that computational methods typically predict neither the severity nor the direction of the effect (increase or decrease of function/affinity). It is true that today's prediction methods cannot directly distinguish between variants that increase and those that decrease binding. Instead, both tend to be predicted as effects. Nevertheless, prediction scores (i.e., the signal strength) on average correlate with the severity of the effect [24]. The concept of “importance for function” never entered the data set choice or development phase when creating SNAP. Still, when applied for residues in ASEdb that the method had never “seen” before, it correctly identified over 70% of the functionally important sites and correctly predicted many to-alanine variants (up to 84%, depending on cutoff) [21].

Comprehensive *in silico* mutagenesis helps studying disease-related proteins

A detailed study of the human melanocortin 4 receptor (hMC4R) demonstrated the value of studying the mutability landscape *in silico* [20]. hMC4R is related to diabetes and to weight regulation. Mutations in hMC4R have been shown to account for approximately 3% of all severe obesity cases (body

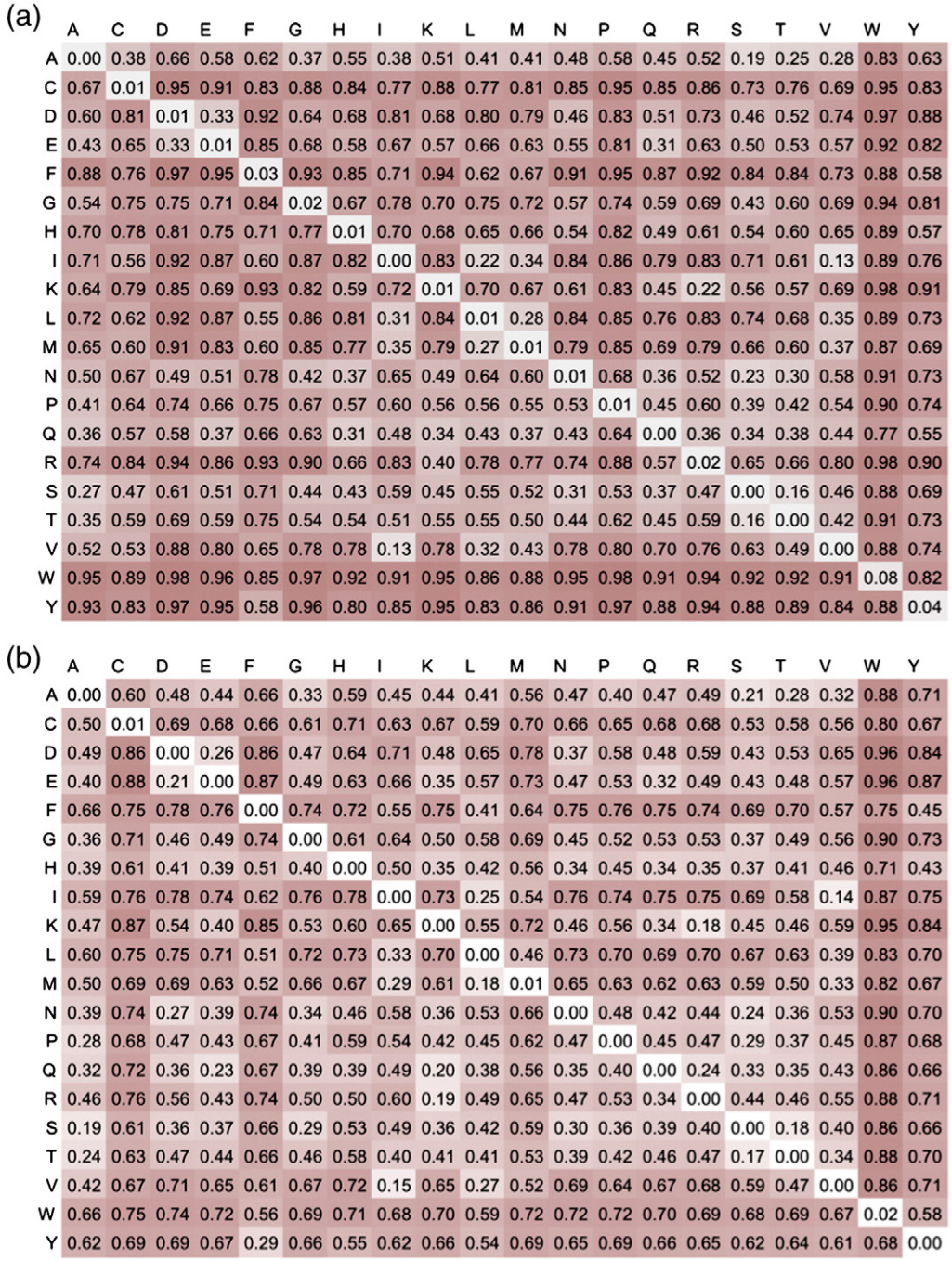

Fig. 3. Effect of pairwise amino acid substitutions in the human exome. Shown is the fraction of substitutions predicted to have an effect for every substitution of every amino acid (*y*-axis) by any other (*x*-axis) in the entire human exome. Results were obtained by locally calculating the predictions for (a) SNAP and (b) SIFT for every possible SAAS in every reviewed protein in the Swiss-Prot database [12] with human origin. Cells are colored according to the fraction of deleterious predictions with high values in red and low values in white. For every prediction for substitutions of amino acid "m" (*y*-axis) by "n" (*x*-axis), we applied the default threshold for each method (SNAP, 0; SIFT, 0.05).

mass index, >40), and consequently, they are the most frequent cause of monogenic obesity in humans [87,88]. MC4R, a member of the G-protein-coupled receptor (GPCR) family, is an integral membrane protein that crosses the lipid bilayer with seven transmembrane helices. SNAP assessed the functional essentiality for each of the 332 residues in hMC4R and the functional impact of all possible variants; predictions were compared to all available experimental data. The predictions of variants with

functional effect and predictions of important regions in hMC4R largely agreed with experimental evidence [20]. Toward this end, we down-weighted mutations expected to be neutral for structure (e.g., hydrophobic to hydrophobic in membrane regions).

Despite this scoring, the computational mutagenesis predicted as many as 118 residues to be functionally important. This seems a substantial over-prediction. Indeed, so far, we have experimental evidence for only 18 residues to be important for function; 15 of these 18 were in the set of 118 residues predicted to have strong impact [20], which is not an impressive performance but much higher than the random 6 in 18. The nsSNP database of effects (SNPdbe [89]) provides experimental links to obesity for 27 residues, 17 of those were in the 118. Only one single residue is found in both sets. Thus, *in silico* mutagenesis correctly predicted 31 of the known 44 positions reported to influence function if mutated.

What about the 74 residues with predictions but without observation (118 predicted, 44 so far experimentally known)? At this point, 74 mutations constitute a relatively large number of high-effect predictions, which cannot be verified due to lack of data. Re-evaluating the predictions, we might apply a more stringent threshold to consider an effect important. For instance, at a threshold with an expected accuracy >95%, 22 residues are predicted to impact function; 10 of those correspond to experimentally known sites, 1 corresponds to a site implicated with obesity and 11 remain without experimental annotations. These might constitute ideal starting points for designing new experiments [20].

Detailed analysis of mutability landscape for a GPCR, the beta-2-adrenergic receptor

To visualize the results of such an *in silico* mutagenesis, we applied SNAP2 (M.H., unpublished results) to another GPCR, the beta-2 adrenergic receptor for which experimental high-resolution 3D structures are available in the Protein Data Bank (PDB) [90] (PDB ID: 2RH1 [15], Fig. 1d; PDB ID: 3PDS [13], Fig. 1e). The predicted high-effect regions cluster around the binding sites and are significantly more abundant on the inside (facing the binding sites) than elsewhere. Strong effects (SNAP > 60; note: score ranges [−100,+100]) are predicted for 57 residues (Fig. 1d, red highlighting) including the two Swiss-Prot annotated [12,91] binding sites D113 and T118. Nine more sites with functional effect annotation in SNPdbe [89] were found. Among these, we find (1) D79, for which a mutation to N was shown to affect binding of catecholamines and to produce an uncoupling between the receptor and stimulatory G-proteins [92,93], and (2) D130, for which mutations to A or N were shown to increase pindolol-stimulated cAMP accumulation [94,95]. Although located in the cytoplasmic region, strong signals also highlight (1) Y141, for which a substitution by F is known to abolish insulin-induced tyrosine phosphorylation and insulin-induced receptor super sensitization [96], and (2) C341, for which a mutation to G was shown to alter binding (uncoupling of receptor) [97].

Thus, 46 sites (57 predicted, 11 experimentally observed) remain with strong effect predictions for which no variants have been tested experimentally. Again we observe a rather large discrepancy between observed and predicted “sensitive to mutation” positions. Some of these predictions will likely just be false positives. However, due to being located in the protein core, others may in fact affect function by structural alterations/misfolding. Application of an even more stringent threshold (score > 80 at >95% expected accuracy) weeds out 38 of these, leaving 12 residues with very strong effect predictions and without current variant annotations. We studied these also in light of EVfold [18,19] (prediction of inter-residue contacts through correlated mutations). Only 2 of the 12 had residue couplings in the realm of the top 5%, namely, W99 and Y199 (Fig. 1e). We could not find any experimental annotation about these two. However, a visual inspection (Fig. 1e) of a 3D structure with irreversibly bound agonist (PDB ID: 3PDS [13]) appears to suggest the two as reasonable targets for experimental verification.

This detailed view of the beta-2-adrenergic receptor provides another example for how useful it might be to analyze the mutability landscape through a complete *in silico* mutagenesis; it highlights functionally important regions and may help in experiment design to probe function locally or test entire regions for docking and drug development. The example also suggests that variant effect prediction might benefit from including inter-residue contact/evolutionary coupling predictions.

Perspective

Comprehensive mutagenesis experiments have furthered our understanding of protein function and continue to provide insight into the mechanisms of pathogenicity and adaption. Novel methodologies and technical advancements reduce the cost of experimental mutagenesis and enable research that was previously impossible. Still, studying the cooperative behavior of amino acids and the combined effect of mutations will remain a laborious and costly task. This is where computational methods are useful to predict the effects of variants upon protein function, structure and pathogenicity. These methods have grown in accuracy both in predicting functional effect (Fig. 2b) and disease-causing

mutations. Can they reach the next level? Can they be used to study the mutability landscape of a protein, that is, to unravel the effects of all possible variants? Here, we argue that the study of such a mutability landscape provides immensely important value and that currently neither experimental nor computational methods completely mine the potential of studying this landscape. Experimental methods remain constrained by the substantial amount of resources such studies would consume. Computational methods remain constrained by the degree to which we can interpret their results. At this point, lack of comprehensive experimental data seems a crucial problem for the development of better computational tools. However, *in silico* analyses of mutability landscape already help to design experiments and are crucial for the intelligent interpretation of deep sequencing/next generation sequencing data.

Acknowledgements

Thanks to Tim Karl, Guy Yachdav and Laszlo Kajan (Technische Universität München) for invaluable help with hardware and software; to Marlena Drabik (Technische Universität München) for administrative support; and to Thomas Hopf and Laszlo Kajan (both Technische Universität München) for helpful discussions and help with the manuscript. Thanks to the developers of PyMOL [98] (Fig. 1d and e) and Matrix2png [99] (Fig. 1b and c) for providing great tools. This work was supported by a grant from the Alexander von Humboldt Foundation through the German Ministry for Research and Education (Bundesministerium fuer Bildung und Forschung). Last, not the least, thanks to all those who deposit their experimental data in public databases and to those who maintain these databases.

Supplementary Data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.jmb.2013.07.028

Received 30 April 2013;
Received in revised form 8 July 2013;
Accepted 19 July 2013
Available online 26 July 2013

Keywords:
complete single mutagenesis;
alanine scanning;
in silico mutagenesis;
exome-wide mutagenesis;
SNP effects

Abbreviations used:
3D, three-dimensional; hMC4R, human melanocortin 4 receptor; GPCR, G-protein-coupled receptor; nsSNP, non-synonymous SNP; PDB, Protein Data Bank; SAAS, single-amino-acid substitution; SIFT, sorting intolerant from tolerant; SNAP, screening for non-acceptable polymorphisms; SNP, single nucleotide polymorphism.

References

[1] Kimura M. Evolutionary rate at the molecular level. Nature 1968;217:624–6.
[2] Ohta T. Near-neutrality in evolution of genes and gene regulation. Proc Natl Acad Sci USA 2002;99:16134–7.
[3] Dunham I, Kundaje A, Aldred SF, Collins PJ, Davis CA, Doyle F, et al. An integrated encyclopedia of DNA elements in the human genome. Nature 2012;489:57–74.
[4] Frazer KA, Murray SS, Schork NJ, Topol EJ. Human genetic variation and its contribution to complex traits. Nat Rev Genet 2009;10:241–51.
[5] Tennessen JA, Bigham AW, O'Connor TD, Fu W, Kenny EE, Gravel S, et al. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. Science 2012;337:64–9.
[6] O'Roak BJ, Vives L, Girirajan S, Karakoc E, Krumm N, Coe BP, et al. Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. Nature 2012;485:246–50.
[7] Rauch A, Wieczorek D, Graf E, Wieland T, Endele S, Schwarzmayr T, et al. Range of genetic mutations associated with severe non-syndromic sporadic intellectual disability: an exome sequencing study. Lancet 2012;380:1674–82.
[8] Thusberg J, Vihinen M. Pathogenic or not? And if so, then how? Studying the effects of missense mutations using bioinformatics methods. Hum Mutat 2009;30:703–14.
[9] McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JP, et al. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nat Rev Genet 2008;9:356–69.
[10] Kiezun A, Garimella K, Do R, Stitziel NO, Neale BM, McLaren PJ, et al. Exome sequencing and the genetic basis of complex traits. Nat Genet 2012;44:623–30.
[11] Samani NJ, Erdmann J, Hall AS, Hengstenberg C, Mangino M, Mayer B, et al. Genomewide association analysis of coronary artery disease. N Engl J Med 2007;357:443–53.
[12] Bairoch A, Apweiler R. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. Nucleic Acids Res 2000;28:45–8.
[13] Rosenbaum DM, Zhang C, Lyons JA, Holl R, Aragao D, Arlow DH, et al. Structure and function of an irreversible agonist-beta(2) adrenoceptor complex. Nature 2011;469:236–40.
[14] Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen bonded and geometrical features. Biopolymers 1983;22:2577–637.

[15] Cherezov V, Rosenbaum DM, Hanson MA, Rasmussen SG, Thian FS, Kobilka TS, et al. High-resolution crystal structure of an engineered human beta2-adrenergic G protein-coupled receptor. Science 2007;318:1258–65.
[16] Ng PC, Henikoff JG, Henikoff S. PHAT: a transmembrane-specific substitution matrix. Predicted hydrophobic and transmembrane. Bioinformatics 2000;16:760–6.
[17] Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci USA 1992;89: 10915–9.
[18] Marks DS, Hopf TA, Sander C. Protein structure prediction from sequence variation. Nat Biotechnol 2012;30: 1072–80.
[19] Hopf TA, Colwell LJ, Sheridan R, Rost B, Sander C, Marks DS. Three-dimensional structures of membrane proteins from genomic sequencing. Cell 2012;149:1607–21.
[20] Bromberg Y, Overton J, Vaisse C, Leibel RL, Rost B. *In silico* mutagenesis: a case study of the melanocortin 4 receptor. FASEB J 2009;23:3059–69.
[21] Bromberg Y, Rost B. Comprehensive *in silico* mutagenesis highlights functionally important residues in proteins. Bioinformatics 2008;24:i207–12.
[22] Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat Protoc 2009;4:1073–81.
[23] Ng PC, Henikoff S. SIFT: predicting amino acid changes that affect protein function. Nucleic Acids Res 2003;31: 3812–4.
[24] Bromberg Y, Rost B. SNAP: predict effect of non-synonymous polymorphisms on function. Nucleic Acids Res 2007;35:3823–35.
[25] Bromberg Y, Yachdav G, Rost B. SNAP predicts effect of mutations on protein function. Bioinformatics 2008;24: 2397–8.
[26] Li B, Krishnan VG, Mort ME, Xin F, Kamati KK, Cooper DN, et al. Automated inference of molecular mechanisms of disease from amino acid substitutions. Bioinformatics 2009;25:2744–50.
[27] Calabrese R, Capriotti E, Fariselli P, Martelli PL, Casadio R. Functional annotations improve the predictive score of human disease-related mutations in proteins. Hum Mutat 2009;30:1237–44.
[28] Reva B, Antipin Y, Sander C. Predicting the functional impact of protein mutations: application to cancer genomics. Nucleic Acids Res 2011;39:e118.
[29] Schwarz JM, Rödelsperger C, Schuelke M, Seelow D. MutationTaster evaluates disease-causing potential of sequence alterations. Nat Methods 2010;7:575–6.
[30] Schaefer C, Rost B. Predict impact of single amino acid change upon protein structure. BMC Genomics 2012;13: S4.
[31] Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, et al. A method and server for predicting damaging missense mutations. Nat Methods 2010;7:248–9.
[32] Olatubosun A, Valiaho J, Harkonen J, Thusberg J, Vihinen M. PON-P: integrated predictor for pathogenicity of missense variants. Hum Mutat 2012;33:1166–74.
[33] Stitziel NO, Kiezun A, Sunyaev S. Computational and statistical approaches to analyzing variants identified by exome sequencing. Genome Biol 2011;12:227.
[34] Cline MS, Karchin R. Using bioinformatics to predict the functional impact of SNVs. Bioinformatics 2011;27: 441–8.
[35] Mah JT, Low ES, Lee E. *In silico* SNP analysis and bioinformatics tools: a review of the state of the art to aid drug discovery. Drug Discovery Today 2011;16:800–9.
[36] Oetting WS. Exploring the functional consequences of genomic variation: the 2010 Human Genome Variation Society Scientific Meeting. Hum Mutat 2011;32:486–90.
[37] Thusberg J, Olatubosun A, Vihinen M. Performance of mutation pathogenicity prediction methods on missense variants. Hum Mutat 2011;32:358–68.
[38] Lesk AM, Levitt M, Chothia C. Alignment of the amino acid sequences of distantly related proteins using variable gap penalties. Protein Eng 1986;1:77–8.
[39] Rost B, Liu J, Nair R, Wrzeszczynski KO, Ofran Y. Automatic prediction of protein function. Cell Mol Life Sci 2003;60: 2637–50.
[40] Rost B, O'Donoghue S, Sander C (1998). Midnight zone of protein structure evolution. EMBL Heidelberg.
[41] Lichtarge O, Bourne HR, Cohen FE. Evolutionarily conserved Galphabetagamma binding surfaces support a model of the G protein-receptor complex. Proc Natl Acad Sci 1996;93: 7507–11.
[42] Sankararaman S, Sjolander K. INTREPID—INformation-theoretic TREe traversal for Protein functional site IDentification. Bioinformatics 2008;24:2445–52.
[43] Sankararaman S, Sha F, Kirsch JF, Jordan MI, Sjolander K. Active site prediction using evolutionary and structural information. Bioinformatics 2010;26:617–24.
[44] Henikoff S, Comai L. Single-nucleotide mutations for plant functional genomics. Annu Rev Plant Biol 2003;54: 375–401.
[45] Zimprich A. Genetics of Parkinson's disease and essential tremor. Curr Opin Neurol 2011;24:318–23.
[46] Zimprich A, Benet-Pages A, Struhal W, Graf E, Eck SH, Offman MN, et al. A mutation in VPS35, encoding a subunit of the retromer complex, causes late-onset Parkinson disease. Am J Hum Genet 2011;89:168–75.
[47] Amberger J, Bocchini CA, Scott AF, Hamosh A. McKusick's Online Mendelian Inheritance in Man (OMIM). Nucleic Acids Res 2009;37:D793–6.
[48] Schaefer C, Bromberg Y, Achten D, Rost B. Disease-related mutations predicted to impact protein function. BMC Genomics 2012;13:S11.
[49] Eisenberg D, Marcotte EM, Xenarios I, Yeates TO. Protein function in the post-genomic era. Nature 2000;405:823–6.
[50] Bogan AA, Thorn KS. Anatomy of hot spots in protein interfaces. J Mol Biol 1998;280:1–9.
[51] Clackson T, Wells JA. A hot spot of binding energy in a hormone-receptor interface. Science 1995;267:383–6.
[52] Kortemme T, Baker D. A simple physical model for binding energy hot spots in protein–protein complexes. Proc Natl Acad Sci USA 2002;99:14116–21.
[53] Ofran Y, Rost B. Protein–protein interaction hot spots carved into sequences. PLoS Comput Biol 2007;3:e119.
[54] DeLano WL, Ultsch MH, de Vos AM, Wells JA. Convergent solutions to binding at a protein–protein interface. Science 2000;287:1279–83.
[55] Konishi S, Iwaki S, Kimura-Someya T, Yamaguchi A. Cysteine-scanning mutagenesis around transmembrane segment VI of Tn10-encoded metal-tetracycline/H(+) antiporter. FEBS Lett 1999;461:315–8.
[56] Qin L, Cai S, Zhu Y, Inouye M. Cysteine-scanning analysis of the dimerization domain of EnvZ, an osmosensing histidine kinase. J Bacteriol 2003;185: 3429–35.

[57] Gardsvoll H, Gilquin B, Le Du MH, Menez A, Jorgensen TJ, Ploug M. Characterization of the functional epitope on the urokinase receptor: complete alanine scanning mutagenesis supplemented by chemical cross-linking. J Biol Chem 2006;281:19260–72.
[58] Thorn KS, Bogan AA. ASEdb: a database of alanine mutations and their effects on the free energy of binding in protein interactions. Bioinformatics 2001;17:284–5.
[59] Markiewicz P, Kleina LG, Cruz C, Ehret S, Miller JH. 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 1994;240:421–33.
[60] Loeb DD, Swanstrom R, Everitt L, Manchester M, Stamper SE, Hutchison CA. Complete mutagenesis of the HIV-1 protease. Nature 1989;340:397–400.
[61] Rennell D, Bouvier SE, Hardy LW, Poteete AR. Systematic mutation of bacteriophage T4 lysozyme. J Mol Biol 1991;222:67–88.
[62] Gottesman ME, Yarmolinsky MB. Integration-negative mutants of bacteriophage lambda. J Mol Biol 1968;31:487–505.
[63] Gottesman S, Gottesman ME. Elements involved in site-specific recombination in bacteriophage lambda. J Mol Biol 1975;91:489–99.
[64] Epstein CJ. Role of the amino acid "code" and of selection for conformation in the evolution of proteins. Nature 1966;210:25–8.
[65] Woese CR, Dugre DH, Dugre SA, Kondo M, Saxinger W. On the fundamental nature and evolution of the genetic code. Cold Spring Harbor Symp Quant Biol 1966;31:723–36.
[66] Dayhoff MO, Barker WC, Hunt LT. Establishing homologies in protein sequences. Method Enzymol 1983;91:524–45.
[67] Zvelebil MJ, Barton GJ, Taylor WR, Sternberg MJE. Prediction of protein secondary structure and active sites using alignment of homologous sequences. J Mol Biol 1987;195:957–61.
[68] Rost B, Sander C. Prediction of protein secondary structure at better than 70% accuracy. J Mol Biol 1993;232:584–99.
[69] Rost B. PHD: predicting one-dimensional protein structure by profile based neural networks. Methods Enzymol 1996;266:525–39.
[70] Rost B, Sander C. Bridging the protein sequence-structure gap by structure predictions. Annu Rev Biophys Biomol Struct 1996;25:113–36.
[71] Altschuh D, Lesk AM, Bloomer AC, Klug A. Correlation of co-ordinated amino acid substitutions with function in viruses related to tobacco mosaic virus. J Mol Biol 1987;193:693–707.
[72] Goebel U, Sander C, Schneider R, Valencia A. Correlated mutations and residue contacts in proteins. Proteins Struct Funct Genet 1994;18:309–17.
[73] Altschuh D, Vernet T, Berti P, Moras D, Nagai K. Coordinated amino acid changes in homologous protein families. Protein Eng 1988;2:193–9.
[74] Pazos F, Ranea JA, Juan D, Sternberg MJ. Assessing protein co-evolution in the context of the tree of life assists in the prediction of the interactome. J Mol Biol 2005;352:1002–15.
[75] Pazos F, Valencia A. *In silico* two-hybrid systemfor the selection of physically interacting protein pairs. Proteins Struct Funct Genet 2002;47:219–27.
[76] de Juan D, Pazos F, Valencia A. Emerging methods in protein co-evolution. Nat Rev Genet 2013;14:249–61.
[77] Marks DS, Colwell LJ, Sheridan R, Hopf TA, Pagnani A, Zecchina R, et al. Protein 3D structure computed from evolutionary sequence variation. PLoS One 2011;6:e28766.
[78] Kowarsch A, Fuchs A, Frishman D, Pagel P. Correlated mutations: a hallmark of phenotypic amino acid substitutions. PLoS Comput Biol 2010;6:e1000923.
[79] Socolich M, Lockless SW, Russ WP, Lee H, Gardner KH, Ranganathan R. Evolutionary information for specifying a protein fold. Nature 2005;437:512–8.
[80] Russ WP, Lowery DM, Mishra P, Yaffe MB, Ranganathan R. Natural-like function in artificial WW domains. Nature 2005;437:579–83.
[81] Halabi N, Rivoire O, Leibler S, Ranganathan R. Protein sectors: evolutionary units of three-dimensional structure. Cell 2009;138:774–86.
[82] McLaughlin RN, Poelwijk FJ, Raman A, Gosal WS, Ranganathan R. The spatial architecture of protein function and adaptation. Nature 2012;491:138–42.
[83] Maxmen A. Exome sequencing deciphers rare diseases. Cell 2011;144:635–7.
[84] Li MX, Gui HS, Kwan JS, Bao SY, Sham PC. A comprehensive framework for prioritizing variants in exome sequencing studies of Mendelian diseases. Nucleic Acids Res 2012;40:e53.
[85] Chun S, Fay JC. Identification of deleterious mutations within three human genomes. Genome Res 2009;19:1553–61.
[86] Liu X, Jian X, Boerwinkle E. dbNSFP: a lightweight database of human nonsynonymous SNPs and their functional predictions. Hum Mutat 2011;32:894–9.
[87] Farooqi IS, Keogh JM, Yeo GS, Lank EJ, Cheetham T, O'Rahilly S. Clinical spectrum of obesity and mutations in the melanocortin 4 receptor gene. N Engl J Med 2003;348:1085–95.
[88] Lubrano-Berthelier C, Le Stunff C, Bougneres P, Vaisse C. A homozygous null mutation delineates the role of the melanocortin-4 receptor in humans. J Clin Endocrinol Metab 2004;89:2028–32.
[89] Schaefer C, Meier A, Rost B, Bromberg Y. SNPdbe: constructing an nsSNP functional impacts database. Bioinformatics 2012;28:601–2.
[90] Bernstein FC, Koetzle TF, Williams GJ, Meyer EF, Brice MD, Rodgers JR, et al. The Protein Data Bank: a computer-based archival file for macromolecular structures. J Mol Biol 1977;112:535–42.
[91] Wu CH, Apweiler R, Bairoch A, Natale DA, Barker WC, Boeckmann B, et al. The Universal Protein Resource (UniProt): an expanding universe of protein information. Nucleic Acids Res 2006;34:D187–91.
[92] Chung FZ, Wang CD, Potter PC, Venter JC, Fraser CM. Site-directed mutagenesis and continuous expression of human beta-adrenergic receptors. Identification of a conserved aspartate residue involved in agonist binding and receptor activation. J Biol Chem 1988;263:4052–5.
[93] Moffett S, Rousseau G, Lagace M, Bouvier M. The palmitoylation state of the beta(2)-adrenergic receptor regulates the synergistic action of cyclic AMP-dependent protein kinase and beta-adrenergic receptor kinase involved in its phosphorylation and desensitization. J Neurochem 2001;76:269–79.

[94] Ballesteros JA, Jensen AD, Liapakis G, Rasmussen SG, Shi L, Gether U, et al. Activation of the beta 2-adrenergic receptor involves disruption of an ionic lock between the cytoplasmic ends of transmembrane segments 3 and 6. J Biol Chem 2001;276:29171–7.

[95] Rasmussen SG, Jensen AD, Liapakis G, Ghanouni P, Javitch JA, Gether U. Mutation of a highly conserved aspartic acid in the beta2 adrenergic receptor: constitutive activation, structural instability, and conformational rearrangement of transmembrane segment 6. Mol Pharmacol 1999;56:175–84.

[96] Valiquette M, Parent S, Loisel TP, Bouvier M. Mutation of tyrosine-141 inhibits insulin-promoted tyrosine phosphorylation and increased responsiveness of the human beta 2-adrenergic receptor. EMBO J 1995;14:5542–9.

[97] O'Dowd BF, Hnatowich M, Caron MG, Lefkowitz RJ, Bouvier M. Palmitoylation of the human beta 2-adrenergic receptor. Mutation of Cys341 in the carboxyl tail leads to an uncoupled nonpalmitoylated form of the receptor. J Biol Chem 1989;264:7564–9.

[98] DeLano WL. The PyMOL molecular graphics system; 2002 .

[99] Pavlidis P, Noble WS. Matrix2png: a utility for visualizing matrix data. Bioinformatics 2003;19:295–6.

[100] Bromberg Y, Kahn PC, Rost B. Neutral and weakly nonneutral sequence variants may define individuality. Proceedings of the National Academy of Sciences; 2003.

Hecht et al. BMC Genomics 2015, 16(Suppl 8):S1
http://www.biomedcentral.com/1471-2164/16/S8/S1

BMC Genomics

RESEARCH Open Access

Better prediction of functional effects for sequence variants

Maximilian Hecht[1*], Yana Bromberg[2,3,4], Burkhard Rost[1,4]

From Varl-SIG 2014: Identification and annotation of genetic variants in the context of structure, function and disease
Boston, MA, USA. 12 July 2014

Abstract

Elucidating the effects of naturally occurring genetic variation is one of the major challenges for personalized health and personalized medicine. Here, we introduce SNAP2, a novel neural network based classifier that improves over the state-of-the-art in distinguishing between effect and neutral variants. Our method's improved performance results from screening many potentially relevant protein features and from refining our development data sets. Cross-validated on >100k experimentally annotated variants, SNAP2 significantly outperformed other methods, attaining a two-state accuracy (effect/neutral) of 83%. SNAP2 also outperformed combinations of other methods. Performance increased for human variants but much more so for other organisms. Our method's carefully calibrated reliability index informs selection of variants for experimental follow up, with the most strongly predicted half of all effect variants predicted at over 96% accuracy. As expected, the evolutionary information from automatically generated multiple sequence alignments gave the strongest signal for the prediction. However, we also optimized our new method to perform surprisingly well even without alignments. This feature reduces prediction runtime by over two orders of magnitude, enables cross-genome comparisons, and renders our new method as the best solution for the 10-20% of sequence orphans. SNAP2 is available at: https://rostlab.org/services/snap2web

Definitions used: Delta, input feature that results from computing the difference feature scores for native amino acid and feature scores for variant amino acid; nsSNP, non-synoymous SNP; PMD, Protein Mutant Database; SNAP, Screening for non-acceptable polymorphisms; SNP, single nucleotide polymorphism; variant, any amino acid changing sequence variant.

Introduction

Some sequence variations matter, changing native protein function or disease-causing potential, while others do not [1]. The distinction between the variants that change protein function and those that are neutral is one key to making sense of the deluge Next Generation Sequencing (NGS) or Deep Sequencing data. Many methods have been developed that address this challenge, spanning a wide range of goals and applications. Some tools are focused on non-coding regions [2-4]; others focus on coding regions and predict the effects of single amino acid variants (non-synonymous single-nucleotide polymorphisms, nsSNPs, or single amino acid substitutions, SAAS) on aspects such as protein structure [5], stability [6-8], binding affinity [9], and function [10,11]. Some methods focus exclusively on the human genome [12,13] and some aspire to identify disease-causing variants [14-16]. Applications to personalized health are obviously important considerations for the developers of such tools. Generally, today's methods are able to distinguish between a set with 100 disease-causing and another with 100 less impacting variants [17,18]. However, identifying one or several variants in an individual responsible for a certain disease is often beyond our reach. Methods have improved significantly by using more protein and variant annotations, as demonstrated in particular in the advance from PolyPhen [12] to PolyPhen-2 [13]. Despite many advances, good data remains missing, in particular

* Correspondence: hecht@rostlab.org
[1]Department of Bioinformatics & Computational Biology, Technische Universität München, Boltzmannstr. 3, 85748 Garching/Munich, Germany
Full list of author information is available at the end of the article

careful annotations of variant neutrality, partially because it is to difficult to carry out "negative experiments" (absence of change [19]).

The best variant effect prediction methods typically use evolutionary information, and a wide variety of features descriptive of protein function and structure. Performance decreases substantially for proteins without informative multiple alignments. Today few human proteins do not map to well-studied sequence families. However, most fully sequenced organisms, predominantly prokaryotic, contribute a substantial fraction of "orphans" (10-20%) [20].

Today's state-of-the-art prediction methods focus on discerning disease-causing variants from the background variation. They, *e.g.* differentiate between human cancer-causing mutations and common variation. This implicitly disregards many variants with functional effects that are not associated with disease. In contrast, the current version of our SNAP (Screening for Non-Acceptable Polymorphisms) method, SNAP2, does not predict the variant effect as "disease or not" but rather as "change of molecular function or not". Similar to most experimental assays, SNAP2 does not directly connect "molecular change" to "impact on organism"; *i.e.* the goal is not to support statements of the type "this single variant improves survival rate". Also similar to many experimental methods, we avoid distinguishing gain-of-function from loss-of-function variants, as these outcomes are often subjective. For instance, gaining in the ΔΔG of binding does NOT imply a "better molecular function" and even the gain of "molecular function" might decrease survival. Here, we introduced several concepts each of which importantly improved over our previous method, SNAP [11]. SNAP2 outperforms its predecessor in three major aspects: better performance, better predictions without alignments, and many orders of magnitude lower runtime.

Methods

Data sets

The training set for SNAP2 resembled that used for development of the original SNAP [11]. In particular, we used the following mixture: variants from PMD (the Protein Mutant Database [21]), residues differing between enzymes with the same experimentally annotated function according to the enzyme classification commission (EC), retrieved from SWISS-PROT [22,23], variants associated with disease as annotated in OMIM (Online Mendelian Inheritance in Men [24]), and HumVar [25].

PMD. We extracted all amino acid changing variants from the Protein Mutant Database [21] (PMD) and mapped these to their corresponding sequences. PMD annotations with 'no change' ('=') qualification (function equivalent to wild-type) were assigned to the 'neutral' class, while variants with any level of increase ('+', '++', '+++') or decrease ('-', '- -', '- - -') in function were assigned to the 'effect' class. Variants with conflicting functional effect annotations were also classified as 'effect'. This approach identified 51,817 variants (neutral: 13,638, effect: 38,179) in 4,061 proteins.

EC. 74% of the PMD data were 'effect' annotations. We balanced this with evidence for neutral variants from enzyme alignments. Assume independent experiments reveal two enzymes to have the same function, i.e. the same EC number (Enzyme Commission number [26]). If these two proteins are very sequence similar, most variants between them are likely 'neutral' with respect to the EC number. While not always correct, the procedure creates a set heavily enriched in truly 'neutral' variants. To turn this concept into data, we aligned all enzymes with experimentally assigned EC numbers in SWISS-PROT [22] using pairwise BLAST [27]. We retrieved all enzyme pairs with pairwise sequence identity >40% and HSSP-values>0 [28-30]. This yielded 26,840 'neutral' variants in 2,146 proteins [11].

Disease. We extracted 22,858 human disease-associated variants in 3,537 proteins from OMIM [24] and HumVar [25]. All disease-associated variants were classified as 'effect'. For many of these variants the change in protein function has not explicitly been demonstrated. These variants may be not causative but, possibly, in linkage disequilibrium with the actual disease-causing variants. Alternatively, they may be affecting splice-sites and/or regulatory elements in the DNA, finally showing up as amino acid substitutions. Hence, by compiling these into the effect class we may be over-estimating functional changes. However, we previously established that relationships to disease provide much stronger evidence for functional effect of variants than any other experimental evidence [17]. Thus, disease variants are clearly strongly enriched in functional significance.

Protein specific studies. We also included data from comprehensive studies of particular proteins, namely LacI repressor from *Escherichia coli* [31] (4,041 variants) and the HIV-1 protease [32] (336 variants). Variants functionally equivalent to wild-type were considered 'neutral'; all others were deemed 'effect'. These variants were not included in training, overlaps (same variant in one of the sets above and these) were removed.

Evaluation sets. We created three subsets of our data for evaluation/development of SNAP2. First, *PMD* + *EC* + *Disease* were compiled into one comprehensive set termed *ALL* with 101,515 variants (40,478 neutral, 61,037 effect) in 9,744 proteins. We also split the *PMD* data into two subsets: one containing only human mutations (*PMD_HUMAN;* 9,657 variants in 678 human sequences) and one consisting of all others (*PMD_NON-HUMAN*; 42,160 variants in 3,383 sequences).

Cross-validation

We clustered our data such that the sets used for training (optimizing neural network connections), cross-training (picking best method) [33,34], and testing (results reported) were not significantly sequence similar. Toward this end, we all-against-all PSI-BLASTed all proteins in our data sets and recorded all hits with E-values<10^{-3}. Starting with these, we built an undirected graph, where vertices are proteins and edges link vertices to the corresponding BLAST hits. We then clustered all proteins using single linkage clustering; *i.e.* all connected vertices were assigned to the same cluster. This yielded 1,241 clusters of related protein sequences with 1 to 1,941 members. We randomly grouped the clusters into ten subsets of roughly similar size. This approach ascertained that no two proteins between any sets were significantly sequence similarity. Due to extremely varied cluster sizes one of these subsets was nearly three times larger than the others. This imbalance was acceptable since the cross-validation procedure ensured sufficiently more training data than testing data in each rotation. In tenfold cross-validation, we rotated through the subsets using eight for training, one for cross-training and the tenth for testing, such that each subset (and therefore each protein) was used for testing exactly once. As a result no variant, protein sequence, or even close homologue, was ever used simultaneously for training and testing. All performance estimates that we reported were solely based on the testing set.

Prediction method

We applied the different machine learning tools in the WEKA suite [35] to our data with default parameters. Support Vector Machines (SVMs) and Neural Networks performed similarly and slightly better than Decision Trees and Random Forests. Due to runtime efficiency, we decided to proceed with standard neural networks. As in similar applications [11,36], we used two output units: one for 'neutral', the other for 'effect'. All free network parameters were optimized on the training (optimizing connection weights) and cross-training (optimizing number of hidden units, learning rate, and momentum; stop training before over-fitting) sets. Tenfold cross-validation implies training ten networks: which one to use for future applications? Taking the "best" of the ten risks over-training. We avoid this by using all ten networks to predict for new proteins, compiling separate averages for 'neutral' and 'effect' over all ten networks. The final prediction is the difference between these averages that ranges from -100 (strongly predicted 'neutral') to +100 (strongly predicted 'effect').

Input features

Biophysical amino acid features and predicted aspects of protein function and structure help to predict the impact of variants. Not knowing connections between residues (our method does not require the knowledge of 3D structures), we scanned sliding windows of up to 21 consecutive residues around the central variant position. We compiled the original SNAP features: biophysical amino acid properties, explicit sequence, PSIC profiles [37], secondary structure and solvent accessibility [38-40], residue flexibility [41], and SWISS-PROT annotations. Additionally, we introduced new features for SNAP2: amino acid properties as provided by the AAindex database [42], predicted binding residues [43], predicted disordered regions [44], proximity to N- and C-terminus, statistical contact potentials [45], co-evolving positions, residue annotations from Pfam [20] and PROSITE [46], low-complexity regions, and other global features such as secondary structure and solvent accessibility composition (Additional File 1, *Input feature calculation*).

Feature selection

In order to determine the optimal feature combination, we systematically sieved through our feature space using greedy bottom-up feature selection. For the following procedure one of the ten training folds (specific to each network) was kept out so that it had no part in feature selection and parameter optimization at any point. We trained ten networks, using 9 of the 10 data subsets: 8 for training and 1 for cross-testing as described above, using each feature and selecting the highest scoring feature separately for each network (highest AUC, Area Under ROC Curve, in cross-training). In the next round, the selected feature was combined with each of the remaining features to train another round of ten networks and the best performing combination of features was selected - again, for each network separately. We repeated until no additional feature improved performance. We considered different sequence window sizes for each feature independently; *i.e.* each feature could be selected in a window of w = 1,5,9,13,17, or 21 consecutive residues around the observed variant at the center of the window.

We tried to avoid local maxima in training via the following steps: S1: Train with balanced data sets [38,40]. S2: Determine the AUC on the cross-training set after each repetition. Record the step with maximal AUC. S3: Train and determine AUC for the cross-training set at least another ten repetitions from the highest-scoring step. Repeat S2-S3 until no additional improvement is recorded.

We collected all features that improved performance on any of the individual networks into a single combined feature set and trained all networks on this set. In a subsequent backward elimination, we removed all features the removal of which did not alter the average overall prediction accuracy. After determining the final feature space, we optimized the number of hidden nodes, learning rate,

and learning momentum to obtain the best-performing network architecture. As an exhaustive screening of the entire parameter space was not intended, we heuristically selected parameter combinations for optimization: learning rate 0.005-0.1, learning momentum 0.01-0.3, and hidden nodes 10-100. The best-performing architecture for each network, as determined by its performance on the corresponding cross-training set, was chosen for the final method.

Finally, we tested the resulting trained networks (of specific feature space and the network architecture each) against the test sets that were initially kept out of feature selection and parameter optimization. Since the performance on these test sets did not differ significantly from that estimated during the optimization procedure, we concluded that we had not over-fitted the networks to the data.

Predicting effects without alignments

We repeated the above feature selection restricted to global features (features based on the entire protein, such as amino acid and secondary structure compositions), amino acid indices, alignment-free secondary structure predictions, and the biophysical amino acid properties. We explicitly left out evolutionary information. We wanted to add a generic average for 'potential effect'. Toward this end, we used the complete version of SNAP2 to predict effects for all possible variants at each residue position in our entire *ALL* set. From these results, we generated a novel amino acid substitution matrix of effect probabilities [47] which we included as an additional feature in the feature selection. This procedure was aimed at developing a method that can be applied without alignments. The resulting method ($SNAP2_{noali}$) predicts functional effects using only single sequences. Note that our SNAP2 implementation selects the best method given the available information, SNAP2 by default and $SNAP2_{noali}$ for orphans. In the latter case, users are notified about the possibly reduced accuracy of predictions.

Performance measures

We evaluated performance via a variety of measures. For simplicity, we used the following standard annotations: True positives (TP) were correctly predicted experimental 'effect' variants, while false positives (FP) were experimentally 'neutral' substitutions incorrectly predicted to have an effect. True negatives (TN) were correctly predicted neutrals and false negatives (FN) were effect variants incorrectly predicted to be neutral. Here, like everywhere else in computational biology, we accept incorrect estimates originating from the triviality that "not observed" does not always imply "not existing", *i.e.* some of the FP might have an effect that was not experimentally tested.

We calculated accuracy (precision) and coverage (recall) separately for 'effect' (Eqn. 1) and 'neutral' (Eqn. 2) predictions:

$$\begin{aligned} Accuracy_{effect} &= Precision_{effect} = Positive\ predictive\ value = \frac{TP}{TP + FP} \\ Coverage_{effect} &= Recall_{effect} = Sensitivity = \frac{TP}{TP + FN} \end{aligned} \quad (1)$$

$$\begin{aligned} Accuracy_{neutral} &= Precision_{neutral} = Negative\ predictive\ value = \frac{TN}{TN + FN} \\ Coverage_{neutral} &= Recall_{neutral} = Specificity = \frac{TN}{TN + FP} \end{aligned} \quad (2)$$

We used the F-measure (F1-Score; Eqn. 3) to asses 'neutral' and 'effect' variants individually. Combined performance was measured by the overall two-state accuracy (Q2; Eqn. 4) and the Matthews Correlation Coefficient (MCC; Eqn. 5).

$$\begin{aligned} F_{effect} &= 2 \cdot \frac{precision_{effect} \cdot recall_{effect}}{precision_{effect} + recall_{effect}} \\ F_{neutral} &= 2 \cdot \frac{precision_{neutral} \cdot recall_{neutral}}{precision_{neutral} + recall_{neutral}} \end{aligned} \quad (3)$$

$$Q_2 = Accuracy = \frac{TP + TN}{(TP + FP + TN + FN)} \quad (4)$$

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (5)$$

Standard deviation and error for all measures were estimated over n = 1000 bootstrap sets; for each set we randomly selected 50% of all variants from the original test set without replacement. Note that due to over-representation of certain protein families, in our experience, bootstrapping without replacement typically yields error estimates that are more accurate than those with replacement. Standard deviation was calculated as the difference of each test set (x_i) from the overall performance $\langle x \rangle$ (Eqn. 6). Standard error was calculated by dividing σ by the square root of sample size (Eqn. 7).

$$\text{Standard deviation (SD)} = \sqrt{\frac{\sum (x_i - \langle x \rangle)^2}{n}} \quad (6)$$

$$\text{Standard error (SE)} = \frac{SD}{\sqrt{(n-1)}} \quad (7)$$

The reliability index (RI; Eqn. 8) for each prediction was computed by normalizing the difference between the two output nodes (one for 'neutral', the other for 'effect') into integers between 0 (low reliability) and 10 (high reliability):

$$RI = 10 \cdot |int(Output_{effect} - Output_{neutral})| \quad (8)$$

Results

SNAP2 significantly improves predictions

First, we assessed the performance of SNAP2 via cross-validation on the original SNAP data. Here, we observed a performance increase over our original SNAP, originating from novel features used in SNAP2. However, by adding in more and better variant data, we found a further (and significantly higher) improvement in performance over SNAP. Many computational methods predict variant effects. As most of these methods focus on predicting disease-associated variants, assessing their performance on our data is inappropriate. Therefore, we explicitly compared SNAP2 only to widely used methods that explicitly aim at the prediction of functional effects: SIFT [10] and PolyPhen-2 [13]. All estimates for the performance of SNAP2 given in this work are based on full cross-validation testing, *i.e.* on data never used for any step in the development. Note that this is not true for other methods in our comparisons.

On the ALL data set (Methods), SNAP2 outperformed its predecessor SNAP [11], as well as both PolyPhen-2 and SIFT (Figure 1). However, the direct comparison is complicated due to a variety of issues. Firstly, the original SNAP was trained on PMD, suggesting a performance overestimate. Secondly, SIFT scores were normalized and optimized for simple defaults. This is implicitly ignored by showing ROC-curves that provide values for a wide set of thresholds that had been deemed non-optimal by the developers. Thirdly, PolyPhen-2 is optimized on human variants that account for only 25% of our *ALL* data. For these, we over-estimate PolyPhen-2's performance. Although the authors assumed that PolyPhen-2 would perform similarly for other eukaryotes, it might not. To address these complications we compared the methods using additional data sets.

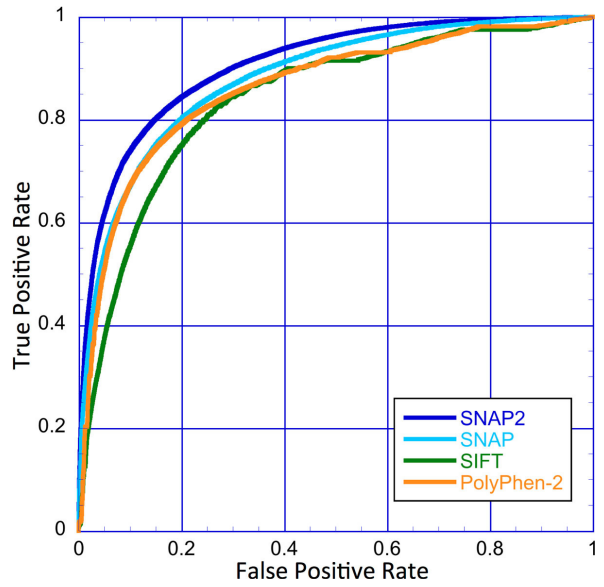

Figure 1 SNAP2 performs best for the ALL data set. This figure shows performance estimates for the *ALL* data set. Our new method SNAP2 (dark blue, AUC = 0.905) outperforms its predecessor SNAP (light blue, AUC = 0.880), PolyPhen-2 (orange, AUC = 0.853) and SIFT (green, AUC = 0.838) over the entire spectrum of the Receiver Operating Characteristic (ROC) curve. Curves are significantly different from each other at a significance level of P < 10-4 as measured by the DeLong method [59]. All SNAP2 results were computed on the test sets not used in training after a rigorous split into training, cross-training and testing. Results for PolyPhen-2 and our original SNAP included some of those proteins in their training, suggesting over-estimated performance.

Performance differed between the human and non-human PMD data

The F-measure for predicting effect (F_{effect}, Eqn. 3), the two state-accuracy (Q2, Eqn. 4), and the Matthew's correlation coefficient (MCC, Eqn. 5) were slightly higher for SNAP2 when tested on the non-human than on the human set (Table 1). For the human PMD data, PolyPhen-2 performed on par with SNAP2, while SIFT was best for predicting neutrals. For the non-human data, SNAP2 was either on par ($F_{neutral}$, Eqn. 3) or outperformed (F_{effect}, Q2, MCC) all other methods (Table 1). Again, this comparison is not entirely fair to SNAP2 and SIFT since the human PMD variants overlapped substantially with the PolyPhen-2 training set, i.e. Table 1 likely over-estimates PolyPhen-2.

Blind method combinations might be worse than a good single method

If in doubt which method is best, users often mix several methods. One strategy is to exclusively consider predictions for which several methods agree. We assessed the benefit of this strategy by applying SNAP2, SIFT and PolyPhen-2 on the *PMD_HUMAN* data set. All methods performed significantly worse for neutral than for effect variants. This can largely be attributed to the difference in the number of variants. The combination of SIFT and PolyPhen-2 improved slightly over SIFT alone for neutral variants (green curve vs. brown arrow/triangle in Figure 2A) and, in terms of accuracy (Eqn. 2) over PolyPhen-2 alone (orange curve vs. brown arrow/triangle in Figure 2A). However, for effect variants combining PolyPhen-2 and SIFT did not improve over the individual methods at all. Moreover, throughout the curves (Figure 2) of both neutral and effect variants, the combined method did not improve over using SNAP2 alone. Methods such as PredictSNP [48], Condel [49], and MetaSNP [50] have been explicitly optimized to combine different methods, mostly to annotate disease-variant relationships (as opposed to functional changes). Such *meta*-methods often tend to improve

Table 1. Method performance on PMD *

	Method	F_{effect} (Eqn. 3)	$F_{neutral}$ (Eqn. 3)	Q2 (Eqn. 4)	MCC (Eqn. 5)
human	*SNAP2*	**78.0% ± 0.6**	46.3% ± 1.3	**68.8% ± 0.7**	0.24 ± 0.01
	PolyPhen-2	**78.4% ± 0.4** **	45.1% ± 1.1 **	**68.9% ± 0.5** **	0.23 ± 0.01 **
	SNAP	74.9% ± 0.5	46.7% ± 1.1	65.8% ± 0.6	0.22 ± 0.01
	SIFT	72.2% ± 0.6	**49.0% ± 1.0**	63.6% ± 0.6	0.23 ± 0.01
non-human	*SNAP2*	**79.9% ± 0.3**	45.8% ± 0.8	**70.7% ± 0.4**	**0.26 ± 0.01**
	PolyPhen-2	77.1% ± 0.4	44.7% ± 0.8	67.6% ± 0.5	0.22 ± 0.01
	SNAP	77.2% ± 0.3	45.5% ± 0.9	67.9% ± 0.5	0.23 ± 0.01
	SIFT	77.0% ± 0.3	45.8% ± 0.8	67.7% ± 0.4	0.23 ± 0.01

* Data set consisting of 9,657 variants (2,788 neutral, 6,869 effect) from 678 human proteins in the top rows and 42,160 variants (10,850 neutral, 31,310 effect) from 3,383 non-human proteins in the bottom rows. For each measure and species group, significantly best results are highlighted in bold. Measures with no bold highlighting indicate absence of a statistically significant best performer.

** Values might over-estimate performance for PolyPhen-2 due to overlap between data set used here and one used for training PolyPhen-2.

over the simple combinations individually attempted by many users and tested here.

SNAP2 is clearly best for difficult cases

Although overall performance levels were similar for all methods tested on the *ALL* data set, the actual predictions for a single variant differed substantially between methods. Variants for which methods agree could be considered "easy" (every method right) or "unsolvable" (no method right). In contrast, variants for which methods disagree could be considered "difficult". This classification yielded 67,912 ***easy*** (~68% of the total; 27,370 neutral and 40,542 effect), 9,624 ***unsolvable*** (~10% of the total; 4,750 neutral and 4,874 effect), and 22,625 ***difficult*** variants (~22% of the total; 7,504 neutral and 15,121 effect). SNAP2 outperformed others on the difficult cases, correctly predicting 69%, as compared to SNAP with 53% and SIFT with 41% compared to 53±1% for random.

We repeated the same analysis for the *PMD_HUMAN* subset (Figure 3). For the 3,963 human variants (1,374 neutral and 2,589 effect) for which any two of the methods disagreed, SNAP2 and PolyPhen-2 were correct in ~58% of the cases compared to 50% for SNAP, 46% for SIFT and 44±1% for random predictions. Again, the PolyPhen-2 training set overlapped with these data, suggesting a performance over-estimate.

In this set of 3,963 human variants, 305 (45 neutral and 260 effect) were only correctly predicted by SNAP2. We investigated these cases in detail, and found that the effect variants in this set often localized to positions at which the variant residue had been observed in another protein in the alignment. For most methods, this implies "neutral" prediction. Indeed, SNAP2noali, the version of

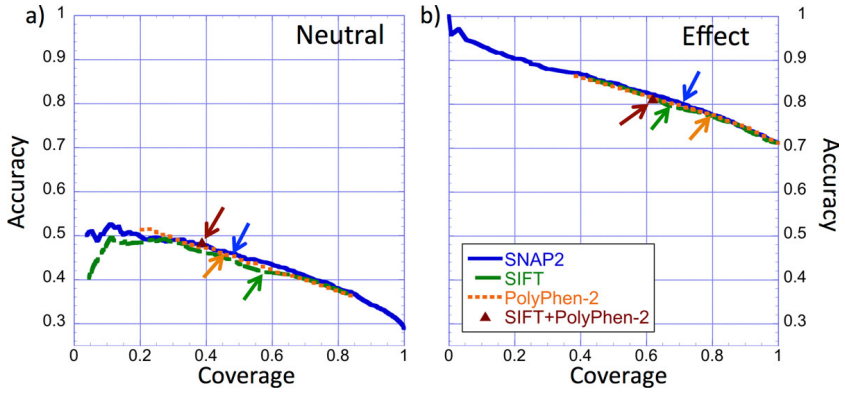

Figure 2 Naïve combination is not better than individual methods for *PMD_HUMAN* data. This figure shows accuracy-coverage curves for the PMD_HUMAN data. The x-axes indicate coverage (also referred to as 'recall'; Eqn. 1.2), *i.e.* the percentage of observed neutral (a) and of observed effect (b) variants that are correctly predicted at the given threshold. The y-axes indicate accuracy (also referred to as 'precision'; Eqn. 1.2), *i.e.* the percentage of neutral (a) and effect (b) variants among all variants predicted in either class at the given threshold. Arrows mark the performance at the default thresholds for our new method SNAP2 (dark blue), for SIFT (green), and for PolyPhen-2 (orange). A brown triangle/arrow marks the performance of a (non-optimized) method that combines PolyPhen-2 and SIFT. This combination did not perform better than SNAP2 alone (brown triangle vs. blue SNAP2 curves).

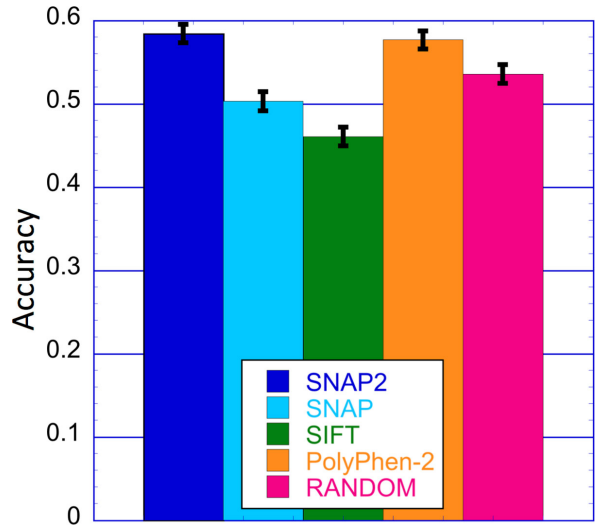

Figure 3 SNAP2 and PolyPhen-2 are best for difficult human variants. Bars mark the two-state accuracy (Q2; Eqn. 4) at the default thresholds for SNAP2 (dark blue), SNAP (light blue), SIFT (green), and PolyPhen-2 (orange). Random prediction performance assuming 60:40 effect:neutral background are given in pink. Analysis is based on 3,963 'difficult' cases (2,589 effect; 1,374 neutral) from *PMD_HUMAN* set. Difficult cases were defined as variants where any of the above method's predictions disagreed; *i.e.* cases where not all methods, excluding random, gave the same prediction.

our method that does not use alignments, predicted 75% of these effect variants at over 90% accuracy, *i.e.* reached a performance substantially above its average for these cases. Thus, one important source of SNAP2 improvement for difficult cases originates from its use on various pieces of information, not just alignments. One example of this improvement is the R109Q variant in the IL4 sequence (interleukin-4 isoform 1 precursor; NCBI reference sequence: NP_000580.1), a pleiotropic cytokine produced by activated T-cells and involved in B-cell activation as well as co-stimulation of DNA synthesis [51]. Variations in this gene were shown to be associated with susceptibility to ischemic stroke [52] and knee osteoarthritis [53]. While our R109Q was not explicitly found to increase disease susceptibility, there is evidence [54] that it reduces T-cell proliferation and receptor binding activity. In this case, the variant glutamine is more conserved in the protein alignment than the human native arginine (11% Q vs. 8% R), making predictions difficult for methods that over-rely on alignments.

Another potential source of improvement, although one for which we could not find explicit and experimentally verified examples in our data, lies in the usage of information about co-evolving residues (Additional File 1, Input Feature Calculation). Specifically, some of the variant positions in this set exhibited (computationally-determined) strong correlations with other positions in the protein, suggesting that this particular feature also made a difference.

Evolutionary information most important, other features vary

The input features related to evolutionary information were consistently most informative for SNAP2 (Additional File 1, Fig. SOM_1: SNAP2 vs. $SNAP2_{noali}$). Which other input features best distinguished neutral from effect depended on the data set. This dependency might originate from annotation inconsistencies and/or set size differences or it might genuinely reflect the data. By selecting the best features separately for subsets of related proteins, we tried to differentiate between these alternatives. The majority of our subsets considered structural features (secondary structure and solvent accessibility) informative, followed by biophysical amino acid properties (more precisely: charge and hydrophobicity). However, the optimal window sizes (number of consecutive residues used as input) for these features differed. For instance, residue flexibility was considered informative by most subsets, but the optimal window size for this feature varied between three and nine residues around the variant.

The final SNAP2 network included the following features: global features (amino acid composition, secondary structure and solvent accessibility composition, and protein length), PSI-BLAST [27] profiles and deltas, PSIC [12] profiles and deltas (differences between mutant and wild-type residue annotations; see Methods for details), residue flexibility, sequence and variant profiles, disorder, secondary structure and relative solvent accessibility and their deltas, physicochemical properties (charge, hydrophobicity, volume, and their deltas), contact potential profiles and deltas, correlated positions and low complexity regions. In addition to these, SWISS-PROT [22] annotations and SIFT [10] predictions were included in SNAP2, if available. For the sequence-only network ($SNAP2_{noali}$) the following features where included: amino acid composition, protein length, sequence and variant profiles, contact potential profiles and delta, volume and hydrophobicity along with the corresponding delta features as well as several amino acid indices from the AAindex [42] (Additional File 1, Table SOM_1).

$SNAP2_{noali}$ important for many proteins

For eight proteins in the *ALL* data set we found fewer than five PSI-BLAST hits in UniProt when we first checked in Oct. 2012. On this tiny set $SNAP2_{noali}$ appeared better than SNAP2 (Eqn. 4: $Q2_{SNAP2noali}$ = 61% vs. $Q2_{SNAP2}$ = 60%; Eqn. 5: $MCC_{SNAP2noali}$ = 0.19 vs. MCC_{SNAP2} = 0.17). PolyPhen-2 made predictions for

only three of these eight proteins (103 variants, $Q2_{Poly-Phen-2}$ = 60%) and SIFT gave no predictions. Recently repeating the analysis, we found homologues for all eight. SNAP2, SIFT and PolyPhen-2 now outperformed $SNAP2_{noali}$. Our "outdated" analysis was important. On the one hand, over 600 human proteins (~3% of all human) still find less than 5 homologues today. On the other hand, for most organisms for which we know the sequences, the corresponding value is much closer to 10-20%, i.e. millions of the proteins we know today can only be handled well by $SNAP2_{noali}$.

For our entire training data, $SNAP2_{noali}$ reached Q2 = 68%, *i.e.* seven percentage points more than for the subset of proteins with small/no families (68% on ALL vs. 61% on NOALI eight protein set). About 10-20% of all proteins in newly sequenced organisms continue not to map anywhere else in today's databases [33,34,55]; for those 10-20% of proteins, $SNAP2_{noali}$ appears to be the best method available to predict the effect of mutations.

Performance confirmed for additional data sets

We avoided over-optimistic performance estimates by removing sequence similarity between proteins used for method development (training/cross-training) and testing. In addition, we also tested our final method on two data sets of variants from the *Escherichia coli* LacI repressor and from the HIV-1 protease (Additional File 1, Table SOM_2). Given the small size and lack of diversity, these results are likely to be more error-prone than our cross-validation estimates. However, they provide independent evidence to estimate the performance of SNAP2: Q2 = 78% for 4,041 LacI variants and Q2 = 72% for 336 HIV-1 variants. None of these variants was used during method development. Moreover, our training data did not contain variants from any homologs of these proteins.

Reliability index allows zooming into best predictions

The difference between the raw output units reasonably estimates prediction confidence [11,36]. We used this difference to define a reliability index (RI, Eqn. 6) and demonstrated its excellent correlation to prediction strength, *i.e.* the reliability index and performance (Figure 4). The final binary predictions (neutral/effect) of SNAP2 are calculated from the network outputs based on the user-defined decision threshold (default: -0.05). By moving the threshold, users can vary the accuracy-coverage balance. Higher thresholds result in more accurate predictions at the cost of covering fewer variants; lower thresholds cover more variants while reducing accuracy. By dialing through the entire threshold spectrum for our non-disease data (PMD/EC data), we estimated and fixed the default decision threshold (Figure 4A). To put this into perspective: when predicting effect/neutral for all variants, SNAP2 is correct in about 75% of its neutral predictions and in 86% of its effect predictions (Figure 4B rightmost points). If users focus on the 50% strongest predictions (Figure 4B; x-axis at 0.5), they could expect the ~92% of the neutral predictions and ~96% of the effect predictions to be correct (RI≥8, Figure 4B). Note that for the purposes of simplified visualization, to display SNAP2 reliability with one digit per residue (*e.g.* to view along with multiple sequence alignments), we projected the actual RI onto integers from 0 (low reliability - worst prediction) to 9 (high reliability - best prediction, Figure 4B).

Discussion

Performance related to experimentally biased balance of neutral vs. effect variants

Machine learning tends to work best when testing and training data are sampled from the same distribution. What are the true data that we want to assess our method upon? One proxy for this type of *truth* might be the next "one million variants" experiment: test 1,000 randomly selected naturally occurring variants in 1,000 representative proteins. One question is: how many variants will be identified as being neutral with respect to protein function? The answer remains importantly vague. Several seemingly contradictory findings are the following. On the one hand, for almost every sequence position (residue) there is a non-native variant that has very little effect on one particular experimental assay [56]. Loosely put: "sequence can change without effect". On the other hand, for almost every residue there is a variant that affects function somehow [56]. Loosely put: "every residue in a protein matters and its variation can change function". There is evidence that individuality of people is partially caused by many slightly non-neutral variants [19]. However, this does not help in estimating the "true" ratio neutral/effect for the next one million. Clearly, today's data sets are strongly biased toward effect variants, simply because it is simpler to measure and easier to publish an effect than a neutral variation. Unfortunately, most of our performance estimates crucially depend on the *true* ratio neutral/effect. Thus, our estimates remain almost as incomplete as the experimental data.

What to expect from variant prediction?

Methods that identify variants related to disease try to pick up changes that are strong enough to cause phenotypic effects that can be classified as disease. This is difficult for two reasons. Firstly, the causality between variant and disease is only clear for the simplest cases such as monogenic or Mendelian diseases. Most diseases appear to be complex, in the sense that they are onset

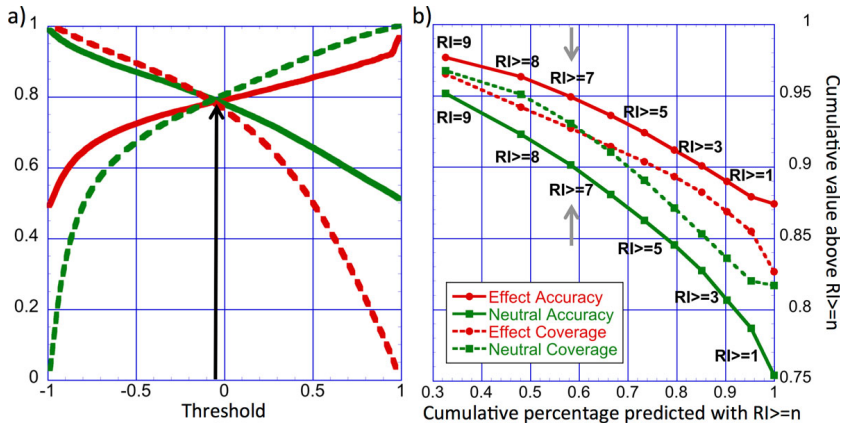

Figure 4 SNAP2 threshold and reliability. The reliability index provides a means of focusing on the most accurate predictions. Panel **(a)** shows SNAP2 performance on the balanced PMD/EC data set over the entire spectrum of accuracy (solid lines) and coverage (dotted lines) for both effect (red) and neutral (green) variants depending on the chosen threshold (x-axis). The default threshold was set to -0.05, where neutral and effect predictions performed alike (black arrow). By moving the decision threshold users can optimize predictive behavior towards their research needs: predictions at higher absolute scores (*e.g.* TP>0.5 or TN<-0.5) are much more likely correct but they are not available for all variants. Panel **(b)** directly relates the reliability index (RI) to the performance on our data. Shown is the cumulative percentage of predictions (x-axis) against accuracy (solid lines) and coverage (dotted lines) above a given reliability index (RI; Methods). Accuracy and coverage are shown separately for neutral (green) and effect (red) predictions. Each marker depicts a reliability threshold ranging from 0 (right most marker, low reliability) to 9 (left most marker, high reliability). Labels for RI >= 2, 4 and, 6 are skipped for simplicity. For instance, 58% of all predictions in our cross-validation were made at reliability levels of 7 or higher (gray arrows). At this reliability, 95% of all effect predictions and 90% of all neutral predictions were correct.

only in the presence of several variants and proper environmental conditions. GWAS have shown that variants associated with disease are found in healthy individuals, and vice versa. Loosely put: the definition of a disease variant may depend on other variants present in the particular genotype of the phenotype carrier. Secondly, even for seemingly clear-cut cases, the classification of "disease" might be misleading. Consider the example of the sickle-cell anemia variants of the hemoglobin B-chain, which can result in a number of chronic health problems on the one hand but grant immunity to some malaria types on the other. In other words, the definition of a disease variant may depend on the environment of the individual.

In contrast to disease, the prediction of the effect of a variant upon molecular function focuses only on the native function of one particular protein. For many examples, such effects are independent of the individual and, often (although not always), of the environment. However, such a focus bears another set of problems: (1) Today's computational methods cannot reliably distinguish between gain and loss of function. They simply predict whether or not the mutation affects native function at all. (2) It is often difficult to relate the strength of a functional effect to its biological relevance. For instance, a "bit" of change in p53 functionality may cause severe phenotypes, whereas a "large" functional effect on other proteins may have little biological impact. In other words, predicted effects have to be put into perspective of the protein in question.

SNAP2 not limited to human variants

Functional effects of sequence variations are not limited to pathogenicity in humans. As most experimental data are human-centric, and as the disease variants are generally most consistent with functional effect [17], SNAP2 performed best for those. This might also explain why for these SNAP2 performed similar to PolyPhen-2 that has been optimized to human data. On non-human variants, however, SNAP2 predictions were most accurate and reliable as compared to other methods. This suggests SNAP2 as a valuable tool for the preliminary analysis of variants in any organism. Specifically, SNAP2 might be the ideal starting point for the comparison of variants between species, *e.g.* human vs. chimp vs. mouse.

Neutral variants predicted worse

All methods performed significantly better for effect than for neutral variants. This in agreement with findings reported in Bromberg *et al* [19] and can be explained in two ways.

(1) The imbalance might originate from incomplete experimental evidence. The effect of variants is typically evaluated on the basis of one or a few phenotypes/assays. If these produce no visible difference as compared to wild-type control the variant is reported as neutral. However, it might still have an effect on other assays that are not performed.

(2) The variants for experimental analysis are usually not selected at random. Instead, researchers prudently

focus on the most important changes; often those changes are related to diseases. Such a prioritized selection samples the feature space incompletely. This may hamper computational detection of relevant patterns for neutral variants. The incomplete sampling may also skew performance estimates: the variants most trivially expected to be neutral might be predicted by the methods but might not be tested experimentally because they are simple to guess. For this reason, comprehensive testing as performed for the *E. Coli* LacI repressor or the HIV-1 protease is an invaluable source of information for computational prediction of variant effects. Such data will likely be crucial in overcoming the neutrality dilemma and will significantly further our understanding of the underlying molecular mechanisms of variant effects.

SNAP2$_{noali}$ succeeded where others failed

We specifically trained a classifier to predict functional effects without using evolutionary information. This unique novel resource might become increasingly useful as ongoing sequencing efforts bring in more data. The current release of the UniRef50 (March 2014) contains ~9.5 million sequence clusters of which over 6.5 million (~68%) contain only one protein, *i.e.* are proteins so far unique to one organism. For those over 6.5 million, very little evolutionary information is available to guide other variant effect predictions and the fraction of orphan clusters appears to be increasing; *i.e.* in October 2012, the UniRef50 contained ~64% orphan clusters - a 4% increase over 1.5 years. This difference might originate from the decreasing quality of increasing sequencing data. However, a similar trend had been observed 12 years ago with arguably more accurate sequencing data [57]. Except for SNAP2$_{noali}$, all methods perform significantly worse for orphans and, in some cases, at the level of throwing a coin. Often they produce no results, which also is at the random level. By including a variety of specific features, we developed a classifier that still achieves a two-state accuracy Q2 around 68% from sequence alone even for these 6.5 million orphan families. This unique type of predicted information might become very relevant for uncharacterized protein families.

Best prediction of difficult cases

By comparing predictions for variants for which commonly applied methods disagreed, we extracted variants that were difficult to classify. For these difficult cases, our new method SNAP2 significantly outperformed SNAP (set ALL-difficult: Q2(snap2) = 69%, Q2(snap) = 53%) and SIFT (Q2(sift) = 41%). For the difficult variants from human PMD, SNAP2 performed just as well as PolyPhen-2, although this comparison gave PolyPhen-2 an unfair advantage because the data set used had partially been used to train PolyPhen-2.

More and better data needed to advance further?

SNAP2 and PolyPhen-2 reached similar levels of performance with rather different approaches, but we made so many so important changes to SNAP that we were surprised not to improve more. Was this because prediction performance has reached a plateau, *i.e.* have we reached the limits for a method using only sequence information as input? Many observations suggest that our data sets remain importantly incomplete. For instance, we observed that our EC data was inconsistent but that we fared worse by leaving it out. We improved a little through the addition of the OMIM data, but possibly only so much so because the data had implicitly already been predicted correctly [17]. In other words: OMIM samples exhibit, on average, extreme signals that are somewhat 'easy' to predict. Thus, adding samples from the top end of the effect distribution did not help improve our prediction of difficult cases where we often find unclear/contradicting signals. Another indication of incompleteness of experimental data was the result that we needed to use all available data to achieve peak performance, *i.e.* smaller subsets reduced performance (data not shown). Still, are we close to a saturation of performance, or can we expect another leap? The lessons learned from advancing secondary structure prediction through the combination of machine learning and evolutionary information suggest that there is yet no way to tell.

Conclusions

We significantly improved over our seven-year-old method SNAP for the prediction of functional effects from single point variants or mutations in the amino acid sequence. SNAP2, the new method improved through more and better data and through more input features. SNAP2 annotates functional effects of variants with little preference to particular species and/or particular types of effects. This allows users to perform bias-free cross-species comparisons, such as looking at sequence positions that differ between human and mouse. We believe that this might be helpful for understanding and predicting disease-causing variation, as well as for facilitating drug development. A measure of prediction reliability (Reliability Index; RI) allows users to focus on the most promising candidates. Additionally, a big achievement of this work is the development of SNAP2$_{noali}$ - a model that predicts effects of variants without using evolutionary information. Ongoing deep-sequencing efforts bring in novel sequences and novel variants alike. Many of these variants occur in sequences without families. Possibly for millions of proteins SNAP2$_{noali}$ provides a reliable prediction of variant

effects and allows for a quick assessment of functionally relevant positions in novel proteins. Both versions of SNAP2 have been optimized towards runtime efficiency to enable large-scale *in silico* mutagenesis studies that probe the landscape of protein mutability [56,58] to learn important news about protein structure and function.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
MH, YB, and BR conceived this work and designed the experiments. MH wrote the software and carried out the experiments. MH and YB collected the data and analyzed the results. MH, YB, and BR wrote, revised, and approved the manuscript.

Acknowledgements
Thanks to Tim Karl, Guy Yachdav, and Laszlo Kajan (TUM) for invaluable help with hardware and software; to Marlena Drabik (TUM) for administrative support; to Peter Hoenigschmid (WZW) and Christian Schaefer (TUM) for helpful discussions; to Shaila C. Roessle (LRZ Munich), Veit Hoehn (TUM), Mark Ofman (TUM), Manfred Roos (TUM), Wiktor Jurkowski (Univ. Luxembourg) and Reinhard Schneider (Univ. Luxembourg) for extensive beta testing of SNAP2. Last, not least, thanks to all those who deposit their experimental data in public databases, and to those who maintain these databases. This work and its publication was supported by a grant from the Alexander von Humboldt foundation through the German Ministry for Research and Education (BMBF: Bundesministerium fuer Bildung und Forschung). This work was also supported by the German Research Foundation (DFG) and the Technische Universität München within the funding programme Open Access Publishing.
This article has been published as part of *BMC Genomics* Volume 16 Supplement 8, 2015: VarI-SIG 2014: Identification and annotation of genetic variants in the context of structure, function and disease. The full contents of the supplement are available online at http://www.biomedcentral.com/bmcgenomics/supplements/16/S8.

Authors' details
[1]Department of Bioinformatics & Computational Biology, Technische Universität München, Boltzmannstr. 3, 85748 Garching/Munich, Germany. [2]Department of Biochemistry and Microbiology, Rutgers University, 76 Lipman Dr, New Brunswick, NJ 08901, USA. [3]Department of Genetics, Rutgers University, 145 Bevier Road, Piscataway, NJ 08854-8082, USA. [4]Institute of Advanced Study (TUM-IAS), Lichtenbergstr. 2a, 85748 Garching/Munich, Germany & WZW - Weihenstephan, Alte Akademie 8, Freising, Germany.

Published: 18 June 2015

References
1. Zuckerkandl E, Pauling L: **Molecules as documents of evolutionary history.** *Journal of Theoretical Biology* 1965, **8**:357-366.
2. Schwarz JM, Rodelsperger C, Schuelke M, Seelow D: **MutationTaster evaluates disease-causing potential of sequence alterations.** *Nat Methods* 2010, **7(8)**:575-576.
3. Cingolani P, Platts A, Wang le L, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM: **A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3.** *Fly* 2012, **6(2)**:80-92.
4. McLaren W, Pritchard B, Rios D, Chen Y, Flicek P, Cunningham F: **Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor.** *Bioinformatics* 2010, **26(16)**:2069-2070.
5. Schaefer C, Rost B: **Predict impact of single amino acid change upon protein structure.** *BMC Genomics* 2012, **13(Suppl 4)**:S4.
6. Dehouck Y, Grosfils A, Folch B, Gilis D, Bogaerts P, Rooman M: **Fast and accurate predictions of protein stability changes upon mutations using statistical potentials and neural networks: PoPMuSiC-2.0.** *Bioinformatics* 2009, **25(19)**:2537-2543.
7. Capriotti E, Fariselli P, Calabrese R, Casadio R: **Predicting protein stability changes from sequences using support vector machines.** *Bioinformatics* 2005, , **21 Suppl 2:** ii54-58.
8. Capriotti E, Fariselli P, Casadio R: **I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure.** *Nucleic Acids Res* 2005, **33(Web Server)**:W306-310.
9. Dehouck Y, Kwasigroch JM, Rooman M, Gilis D: **BeAtMuSiC: Prediction of changes in protein-protein binding affinity on mutations.** *Nucleic Acids Res* 2013, **41(Web Server)**:W333-339.
10. Ng PC, Henikoff S: **SIFT: Predicting amino acid changes that affect protein function.** *Nucleic Acids Res* 2003, **31(13)**:3812-3814.
11. Bromberg Y, Rost B: **SNAP: predict effect of non-synonymous polymorphisms on function.** *Nucleic Acids Res* 2007, **35(11)**:3823-3835.
12. Sunyaev SR, Eisenhaber F, Rodchenkov IV, Eisenhaber B, Tumanyan VG, Kuznetsov EN: **PSIC: profile extraction from sequence alignments with position-specific counts of independent observations.** *Protein Eng* 1999, **12(5)**:387-394.
13. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR: **A method and server for predicting damaging missense mutations.** *Nat Methods* 2010, **7(4)**:248-249.
14. Li B, Krishnan VG, Mort ME, Xin F, Kamati KK, Cooper DN, Mooney SD, Radivojac P: **Automated inference of molecular mechanisms of disease from amino acid substitutions.** *Bioinformatics* 2009, **25(21)**:2744-2750.
15. Calabrese R, Capriotti E, Fariselli P, Martelli PL, Casadio R: **Functional annotations improve the predictive score of human disease-related mutations in proteins.** *Human mutation* 2009, **30(8)**:1237-1244.
16. Reva B, Antipin Y, Sander C: **Predicting the functional impact of protein mutations: application to cancer genomics.** *Nucleic Acids Res* 2011, **39(17)**: e118.
17. Schaefer C, Bromberg Y, Achten D, Rost B: **Disease-related mutations predicted to impact protein function.** *BMC Genomics* 2012, **13(Suppl 4)**:S11.
18. Cline MS, Karchin R: **Using bioinformatics to predict the functional impact of SNVs.** *Bioinformatics* 2011, **27(4)**:441-448.
19. Bromberg Y, Kahn PC, Rost B: **Neutral and weakly nonneutral sequence variants may define individuality.** *Proceedings of the National Academy of Sciences of the United States of America* 2013, **110(35)**:14255-14260.
20. Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C, Pang N, Forslund K, Ceric G, Clements J, *et al*: **The Pfam protein families database.** *Nucleic Acids Res* 2012, **40(Database)**:D290-301.
21. Kawabata T, Ota M, Nishikawa K: **The Protein Mutant Database.** *Nucleic Acids Res* 1999, **27(1)**:355-357.
22. Bairoch A, Apweiler R: **The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000.** *Nucleic Acids Res* 2000, **28(1)**:45-48.
23. Dimmer EC, Huntley RP, Alam-Faruque Y, Sawford T, O'Donovan C, Martin MJ, Bely B, Browne P, Mun Chan W, Eberhardt R, *et al*: **The UniProt-GO Annotation database in 2011.** *Nucleic Acids Res* 2011.
24. Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA: **Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders.** *Nucleic Acids Res* 2005, **33(Database)**: D514-517.
25. Capriotti E, Calabrese R, Casadio R: **Predicting the insurgence of human genetic diseases associated to single point protein mutations with support vector machines and evolutionary information.** *Bioinformatics* 2006, **22(22)**:2729-2734.
26. Webb EC: **Enzyme Nomenclature 1992.** *Recommendations of the Nomenclature committee of the International Union of Biochemistry and Molecular Biology.* 1992 edition. New York: Academic Press;1992.
27. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25(17)**:3389-3402.
28. Sander C, Schneider R: **Database of homology-derived protein structures and the structural meaning of sequence alignment.** *Proteins* 1991, **9(1)**:56-68.
29. Rost B: **Twilight zone of protein sequence alignments.** *Protein Eng* 1999, **12(2)**:85-94.
30. Mika S, Rost B: **UniqueProt: creating representative protein sequence sets.** *Nucleic Acids Res* 2003, **31(13)**:3789-3791.
31. Markiewicz P, Kleina LG, Cruz C, Ehret S, Miller JH: **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* 1994, **240(5)**:421-433.

32. Loeb DD, Swanstrom R, Everitt L, Manchester M, Stamper SE, Hutchison CA III: **Complete mutagenesis of the HIV-1 protease.** *Nature* 1989, **340(6232)**:397-400.
33. Mistry J, Kloppmann E, Rost B, Punta M: **An estimated 5% of new protein structures solved today represent a new Pfam family.** *Acta crystallographica Section D, Biological crystallography* 2013, **69(Pt 11)**:2186-2193.
34. Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C, Pang N, Forslund K, Ceric G, Clements J, *et al*: **The Pfam protein families database.** *Nucleic Acids Res* 2012, **40(Database)**:D290-301.
35. Frank E, Hall M, Trigg L, Holmes G, Witten IH: **Data mining in bioinformatics using Weka.** *Bioinformatics* 2004, **20(15)**:2479-2481.
36. Rost B, Sander C: **Prediction of protein secondary structure at better than 70% accuracy.** *J Mol Biol* 1993, **232**:584-599.
37. Sunyaev SR, Eisenhaber F, Rodchenkov IV, Eisenhaber B, Tumanyan VG, Kuznetsov EN: **PSIC: profile extraction from sequence alignments with position-specific counts of independent observations.** *Protein Engineering* 1999, **12(5)**:387-394.
38. Rost B: **PHD: predicting one-dimensional protein structure by profile based neural networks.** *Methods in Enzymology* 1996, **266**:525-539.
39. Rost B, Sander C: **Conservation and prediction of solvent accessibility in protein families.** *Proteins* 1994, **20(3)**:216-226.
40. Rost B, Sander C: **Prediction of protein secondary structure at better than 70% accuracy.** *J Mol Biol* 1993, **232(2)**:584-599.
41. Schlessinger A, Yachdav G, Rost B: **PROFbval: predict flexible and rigid residues in proteins.** *Bioinformatics* 2006, **22(7)**:891-893.
42. Kawashima S, Kanehisa M: **AAindex: amino acid index database.** *Nucleic Acids Res* 2000, **28(1)**:374.
43. Ofran Y, Rost B: **ISIS: interaction sites identified from sequence.** *Bioinformatics* 2007, **23(2)**:e13-16.
44. Schlessinger A, Punta M, Yachdav G, Kajan L, Rost B: **Improved disorder prediction by combination of orthogonal approaches.** *PLoS One* 2009, **4(2)**:e4433.
45. Simons KT, Ruczinski I, Kooperberg C, Fox BA, Bystroff C, Baker D: **Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins.** *Proteins* 1999, **34(1)**:82-95.
46. Sigrist CJ, Cerutti L, de Castro E, Langendijk-Genevaux PS, Bulliard V, Bairoch A, Hulo N: **PROSITE, a protein domain database for functional characterization and annotation.** *Nucleic Acids Res* 2010, **38(Database)**: D161-166.
47. Hoehn V: **In-depth comparison of predicted high-and low-impact SNPs from the 1,000 Genomes Project.** *Master Thesis* Technische Universität München; 2012.
48. Bendl J, Stourac J, Salanda O, Pavelka A, Wieben ED, Zendulka J, Brezovsky J, Damborsky J: **PredictSNP: robust and accurate consensus classifier for prediction of disease-related mutations.** *PLoS Comput Biol* 2014, **10(1)**:e1003440.
49. Gonzalez-Perez A, Lopez-Bigas N: **Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel.** *American journal of human genetics* 2011, **88(4)**:440-449.
50. Capriotti E, Altman RB, Bromberg Y: **Collective judgment predicts disease-associated single nucleotide variants.** *BMC Genomics* 2013, **14(Suppl 3)**:S2.
51. Yokota T, Otsuka T, Mosmann T, Banchereau J, DeFrance T, Blanchard D, De Vries JE, Lee F, Arai K: **Isolation and characterization of a human interleukin cDNA clone, homologous to mouse B-cell stimulatory factor 1, that expresses B-cell-and T-cell-stimulating activities.** *Proceedings of the National Academy of Sciences of the United States of America* 1986, **83(16)**:5894-5898.
52. Zee RY, Cook NR, Cheng S, Reynolds R, Erlich HA, Lindpaintner K, Ridker PM: **Polymorphism in the P-selectin and interleukin-4 genes as determinants of stroke: a population-based, prospective genetic analysis.** *Human molecular genetics* 2004, **13(4)**:389-396.
53. Yigit S, Inanir A, Tekcan A, Tural E, Ozturk GT, Kismali G, Karakus N: **Significant association of interleukin-4 gene intron 3 VNTR polymorphism with susceptibility to knee osteoarthritis.** *Gene* 2014, **537(1)**:6-9.
54. Ramanathan L, Ingram R, Sullivan L, Greenberg R, Reim R, Trotta PP, Le HV: **Immunochemical mapping of domains in human interleukin 4 recognized by neutralizing monoclonal antibodies.** *Biochemistry* 1993, **32(14)**:3549-3556.
55. Liu J, Rost B: **Comparing function and structure between entire proteomes.** *Protein Science* 2001, **10(10)**:1970-1979.
56. Hecht M, Bromberg Y, Rost B: **News from the protein mutability landscape.** *J Mol Biol* 2013, **425(21)**:3937-3948.
57. Liu J, Rost B: **Comparing function and structure between entire proteomes.** *Protein science : a publication of the Protein Society* 2001, **10(10)**:1970-1979.
58. Bromberg Y, Rost B: **Comprehensive in silico mutagenesis highlights functionally important residues in proteins.** *Bioinformatics* 2008, **24(ECCB Proceedings)**:i207-i212.
59. DeLong ER, DeLong DM, Clarke-Pearson DL: **Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach.** *Biometrics* 1988, **44(3)**:837-845.

doi:10.1186/1471-2164-16-S8-S1
Cite this article as: Hecht *et al.*: Better prediction of functional effects for sequence variants. *BMC Genomics* 2015 **16**(Suppl 8):S1.

Supporting online material for: Better prediction of functional effects for sequence variants

Maximilian Hecht, Yana Bromberg & Burkhard Rost

Table of Contents for Supporting Online Material

Short description of Supporting Online Material

This SOM contains a detailed description of features and their extraction for use in the neural network predictor. We also included three tables (1) listing the cluster representatives from the AAindex database, that were selected as helpful features in $SNAP2_{noali}$, (2) a performance comparison on independent protein-specific data, namely the HIV-1 protease and the *Escherichia Coli* LacI repressor and (3) a table showing performance values on our comprehensive ALL (main manuscript, methods section) data set. Moreover, this SOM includes a figure (Fig. SOM_1) showing the performance of SNAP2 and $SNAP2_{noali}$ in comparison to SIFT and random predictions.

Material

Input feature calculation. In order to use amino acid and protein properties in neural networks these have to be presented as normalized numerical values. The following section describes the exact calculation or extraction of these values.

Delta features. Where applicable, we calculated *delta features* that describe the change in certain features between the native amino acid and its variant. All *delta features* are encoded by two nodes per residue: one for the "severity" (absolute difference between wildtype and mutant value) the other for the "direction" ('1' if positive and '0' if negative) of change.

Biophysical properties. In addition to mass, volume, charge, hydrophobicity and the presence of C-beta branching amino acids (as already present in SNAP) we collected one representative for each cluster of correlated amino acid indices from the AAindex database [1]. These indices are matrices containing values for each amino acid (or pair of amino acids) that cover a variety of amino acid properties and features derived from these (Table SOM_1). We extracted the corresponding (already normalized) value for each residue in the window, resulting in w input values. Then we calculated the two-node delta feature. The first node was the absolute difference between the wildtype and the mutant value.

Binding residues. We used ISIS [2] to predict the protein-protein binding sites and DISIS [3] to predict the protein-DNA binding sites. We extracted both the binary prediction (binding/non-binding) and the raw prediction score for each residue in the window (21 * 2 = 42 input nodes).

Disordered regions. We used the META-Disorder predictor tool (MD; [4]) tool to calculate a three-node disorder feature for all residues in the window: We extracted the binary per-residue prediction (disordered/not-disordered) and the prediction reliability.

Proximity to N- and C-terminus. We calculated the proximity of the variant position to each terminus individually as the normalized number of residues between terminus and the position of interest (2*1 = 2 input nodes).

Contact potentials. We extracted normalized distance-dependent statistical potentials (for contacts within 5 Ångstrøms=0.5nm) [5]. For both native amino acid and variant, we extracted the potential as a 20-node feature. Additionally, we calculated the delta values for this feature (difference between native and variant) for their eight (four residues before and after) sequence neighbors (20*2 + 8*2 = 56 input nodes).

Co-evolving positions. We estimated the co-evolution of positions in a multiple sequence alignment following the approach from [6]. For each position in the multiple alignment we used the OMES [7] algorithm to calculate the correlation

with any other position. The OMES method compares the observed co-occurrence of amino acid X at position i and amino acid Y at position j to the expected co-occurrence at positions i and j. This pairwise comparison yielded a ranking of all positions based on their pairwise correlation to any other position. From these, we extracted a six-node feature indicating the rank and the score (i.e. the deviation from the expectation value) for the three positions most correlated with the mutation position (2*3 = 6 input nodes).

Residue annotation. In addition to SWISS-PROT annotations and SIFT predictions as already used in SNAP we considered residue annotation from Pfam [8] and PROSITE [9] to describe native and variant amino acids: (i) We determined whether the position was part of a PfamA domain. If so, we collected metrics of domain conservation and the posterior probability of native and variant belonging to that domain (4 input nodes). (ii) From PROSITE we extracted a binary single-node feature for all residues in the window indicating whether the specific residue is part of a PROSITE pattern (21 input nodes).

Low-complexity regions. We used the SEG [10] algorithm to mask protein regions with low-complexity. From this masking, we extracted a feature of 21 binary input nodes indicating whether a mutation is in or close to a low-complexity region.

Global features. We added global sequence information by calculating four features: The amino acid composition as the relative frequency of each amino acid (20 amino acids + 1 unknown = 21 input nodes); the sequence length feature encoding the protein length in 6 bins (0-60, 61-120, 121-180,181-240, 241-300, >300; 6 input nodes); the secondary structure composition and the solvent accessibility composition, each as a twelve-node binary feature using four bins (0-25%, 26%-50%, 51%-75%, 76%-100%) for each state: helix-strand-other or buried-intermediate-exposed (2 * 12 = 24 input nodes).

Table SOM_1: Input features selected from AAindex *

AAindex [1] accession	Description
VINM940103	Normalized flexibility parameters (B-values) for each residue surrounded by one rigid neighbour [11]
BLAM930101	Alpha helix propensity [12]
DAYM780201	Relative mutability [13]
QIAN880123	Weights for beta-sheet [14]
KLEP840101	Prediction of protein function from sequence properties; Discriminant analysis of a data base: Net charge [15]
SNEP660101	Relations between chemical structure and biological activity in peptides: Principal component I [16]
RICJ880113	Relative preference values of amino acids at C2 [17]
SIMK990101	Distance-dependent statistical potential (contacts within 0-5 Angstroms) [5]

* We listed the best-performing input features, i.e. amino acid indices that were selected by the feature selection procedure. Other indices from the corresponding clusters performed similarly. For each of these features both window-based and delta features were included into the final sequence-only network $SNAP2_{noali}$.

Table SOM_2: Performance on independent data sets *

Method	LacI repressor	HIV-1 protease
SIFT	72.2% ± 1.0	79.5% ± 3.2
SNAP	72.0% ± 1.0	78.3% ± 3.0
SNAP2	78.3% ± 0.9	74.1% ± 3.2

- Shown is the overall two-state accuracy (Q2 value; Method section) on 4041 LacI mutants and 336 HIV-1 protease mutants for SIFT, SNAP and SNAP2.

Table SOM_3: Performance estimates on ALL data set.

	Q2	F1 (neutral)	F1 (effect)	MCC	ROC AUC
SNAP2	83.5%	0.79	0.87	0.65	0.91
SNAP	80.1%	0.76	0.83	0.59	0.88
SIFT	77.4%	0.74	0.83	0.54	0.84
PolyPhen-2	80.8%	0.75	0.84	0.60	0.85

* Performance estimates were obtained from cross-validation for SNAP2. For all methods the default thresholds were applied. Estimates are based on all variants from our ALL data set (see methods section; Data).

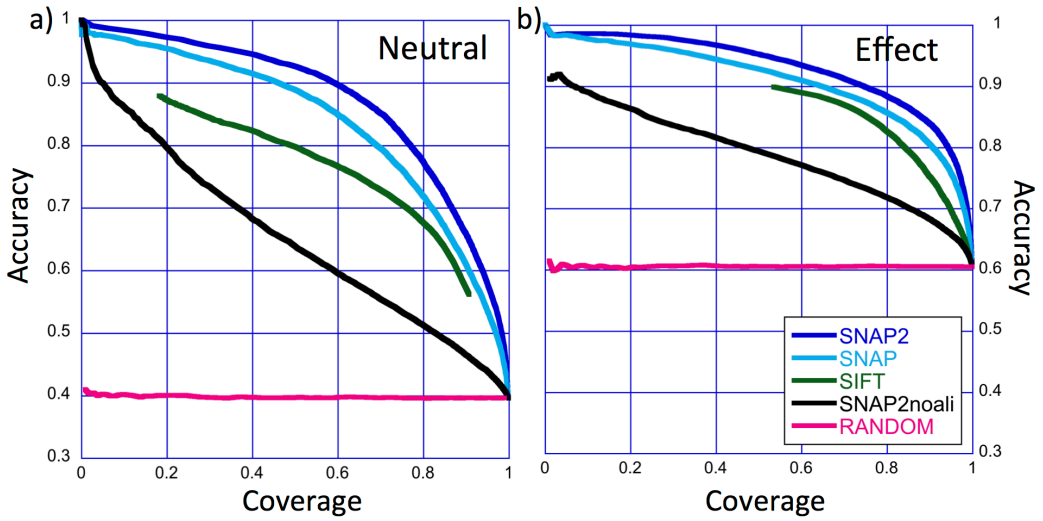

Figure SOM_1: Accuracy-Coverage curves for ALL data. These figures show performance on the *ALL* data set. Our new method SNAP2 (dark blue) outperforms its predecessor (SNAP, light blue), and SIFT (green) for both the variants that do not affect function (neutral, a) and for those that affect function (b). The x-axes indicate coverage/recall (Eqn. 1,2), *i.e.* the percentage of observed neutral (a) and effect (b) variants that are correctly predicted at the given threshold. The y-axes indicate accuracy/precision (Eqn. 1,2), *i.e.* the percentage of neutral (a) and effect (b) variants among all variants predicted in either class at the given threshold. The dark line ($SNAP2_{noali}$) marks the performance of a SNAP2 version that does not use any information from sequence alignment. All results are computed on the test sets not used in training. A pink line marks the performance of a random predictor.

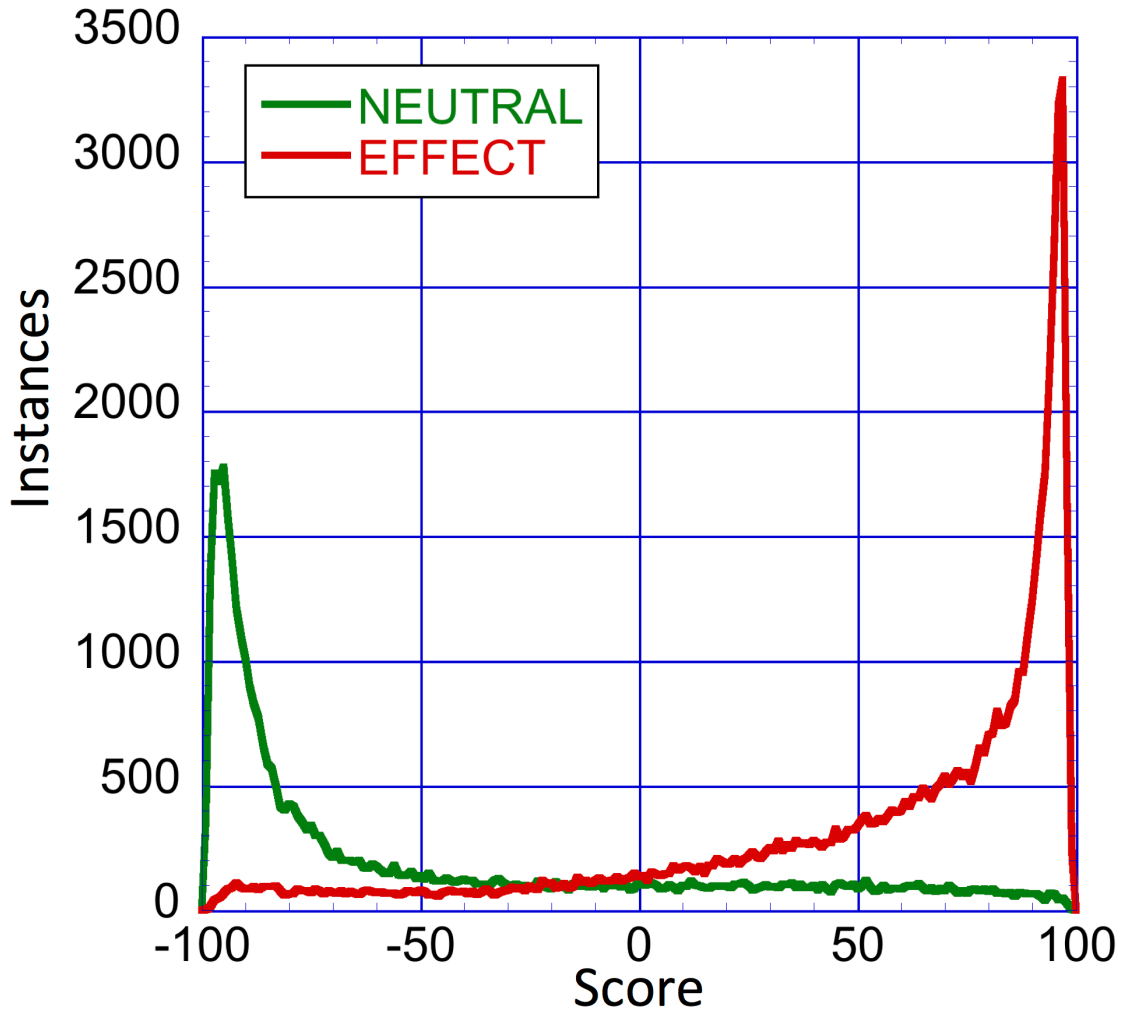

Figure SOM_2: Score distribution for SNAP2 on ALL data. Shown is the number of instance (y-axis) for each score (x-axis). Effect variants (red) mostly have predicted scores > 0 while neutral variants (green) are predominantly predicted at scores < 0.

References for Supporting Online Material

1. Kawashima, S. & Kanehisa, M. (2000). AAindex: amino acid index database. *Nucleic acids research* **28**, 374.

2. Ofran, Y. & Rost, B. (2007). ISIS: interaction sites identified from sequence. *Bioinformatics* **23**, e13-6.

3. Ofran, Y., Mysore, V. & Rost, B. (2007). Prediction of DNA-binding residues from sequence. *Bioinformatics* **23**, i347-53.

4. Schlessinger, A., Punta, M., Yachdav, G., Kajan, L. & Rost, B. (2009). Improved disorder prediction by combination of orthogonal approaches. *PloS one* **4**, e4433.

5. Simons, K. T., Ruczinski, I., Kooperberg, C., Fox, B. A., Bystroff, C. & Baker, D. (1999). Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins. *Proteins* **34**, 82-95.

6. Kowarsch, A., Fuchs, A., Frishman, D. & Pagel, P. (2010). Correlated mutations: a hallmark of phenotypic amino acid substitutions. *PLoS computational biology* **6**.

7. Fodor, A. A. & Aldrich, R. W. (2004). Influence of conservation on calculations of amino acid covariance in multiple sequence alignments. *Proteins* **56**, 211-21.

8. Punta, M., Coggill, P. C., Eberhardt, R. Y., Mistry, J., Tate, J., Boursnell, C., Pang, N., Forslund, K., Ceric, G., Clements, J., Heger, A., Holm, L., Sonnhammer, E. L., Eddy, S. R., Bateman, A. & Finn, R. D. (2012). The Pfam protein families database. *Nucleic acids research* **40**, D290-301.

9. Sigrist, C. J., Cerutti, L., de Castro, E., Langendijk-Genevaux, P. S., Bulliard, V., Bairoch, A. & Hulo, N. (2010). PROSITE, a protein domain database for functional characterization and annotation. *Nucleic acids research* **38**, D161-6.

10. Wootton, J. C. & Federhen, S. (1993). Statistics of local complexity in amino acid sequences and sequence databases. *Computers & chemistry* **17**, 149-163.

11. Vihinen, M., Torkkila, E. & Riikonen, P. (1994). Accuracy of protein flexibility predictions. *Proteins* **19**, 141-9.

12. Blaber, M., Zhang, X. J. & Matthews, B. W. (1993). Structural basis of amino acid alpha helix propensity. *Science* **260**, 1637-40.

13. Dayhoff, M. O., Schwartz, R. M. & Orcutt, B. C. (1978). A model of evolutionary change in proteins. In *Atlas of protein sequence and structure* (Dayhoff, M. O., ed.), Vol. 5. National Biomedical Research Foundation, Washington, DC.

14. Qian, N. & Sejnowski, T. J. (1988). Predicting the secondary structure of globular proteins using neural network models. *Journal of molecular biology* **202**, 865-884.

15. Klein, P., Kanehisa, M. & DeLisi, C. (1984). Prediction of protein function from sequence properties: Discriminant analysis of a data base. *Biochimica et Biophysica Acta (BBA)-Protein Structure and Molecular Enzymology* **787**, 221-226.

16. Sneath, P. (1966). Relations between chemical structure and biological activity in peptides. *Journal of theoretical biology* **12**, 157-195.

17. Richardson, J. S. & Richardson, D. C. (1988). Amino acid preferences for specific locations at the ends of alpha helices. *Science (New York, NY)* **240**, 1648.