TECHNISCHE UNIVERSITÄT MÜNCHEN

Lehrstuhl für Bioinformatik

Predicting the structural effect upon single amino acid exchange

Christian Alexander Schäfer

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. Dmitrij Frischmann

Die Dissertation wurde am 25.09.2012 bei der Technischen Universität München eingereicht und durch die Fakultät für Informatik am 14.12.2012 angenommen.

Contents

Abstract

Through the advent of next generation sequencing methods, it has become feasible to fine-map causative genetic markers to interesting traits in large scale. Of particular interest are point mutations that alter a single amino acid in a protein (non-synonymous single nucleotide polymorphisms, nsSNPs). These single amino acid exchanges potentially affect protein structure or function and could result in genetic diseases. Despite all efforts, the vast majority of nsSNPs lacks experimental verification of their possible disease phenotype.

In this work, we studied structural effects induced by amino acid changes and their implications for protein function and disease. As a first step, we *in-silico* altered amino acids in native protein sequences and monitored coarse-grained consequences on predicted secondary structure and predicted protein disorder. Although our results suggested that secondary structure is an *intrinsic* feature of amino acid sequences, our mutation analysis revealed a highly dynamic picture in the details: predicted helices, strands and short disorder continuously came and went while long disorder completely disappeared in random sequences.

To predict effects in more detail, we proposed a novel structure-centric view of effect. From protein structures, we compiled pairs of pentapeptides, each differing in its two central amino acids. We distinguished pairs of two structurally similar peptides from pairs of two dissimilar ones. This set served to induce a machine-learning model trained on sequence-derived features and subjected to separate structural neutral from effect pairs. Comprehensive validation revealed high predictive performance and suggested that local structural change upon single amino acid change can be predicted from protein sequence.

We predicted structural and functional change in an extensive compilation of effect and disease annotated nsSNPs. Our findings showed that, first, observed effects in protein function, stability and disease were enriched in mutations with strong predicted structural effect. Second, disease-related mutants displayed strong predicted functional effect. This indicated that both effects raise the likelihood for disease.

Motivated by these results, we combined predictions of structural and functional effect to separate disease-related from neutral variants. Based on this analysis we concluded that predicted functional impact alone sufficed to accurately predict whether a nsSNP leads to disease or not.

List of publications and conferences

The work at hand constitutes a cumulative dissertation. The methodologies and results as presented here – in particular sections 2.1-2.5 and 3.1-3.5 – have been published in the following peer-reviewed articles. The manuscripts have been appended to this dissertation.

- **Christian Schaefer**, Avner Schlessinger, Burkhard Rost (2010). *Protein secondary structure appears to be robust under in silico evolution while protein disorder appears not to be.* Bioinformatics, 26(5):625-631.
- **Christian Schaefer**, Alice Meier, Burkhard Rost, Yana Bromberg (2012). *SNPdbe: Constructing an nsSNP functional impacts database.* Bioinformatics, 28(4):601-2.
- **Christian Schaefer**, Burkhard Rost (2012). *Predict impact of single amino acid change upon protein structure.* BMC Genomics, 13(Suppl 4):S4.
- **Christian Schaefer**, Yana Bromberg, Dominik Achten, Burkhard Rost (2012). *Disease-related mutations predicted to impact protein function.* BMC Genomics, 13(Suppl 4):S11.

The contents of the following peer-reviewed publications and conference attendances have not been used directly for the work at hand.

- Avner Schlessinger, **Christian Schaefer**, Esmeralda Vicedo, Markus Schmidberger, Marco Punta, Burkhard Rost (2011). *Protein disorder–a breakthrough invention of evolution?* Current Opinion in Structural Biology, 21(3): 412–418.
- **Christian Schaefer**, Burkhard Rost. *Can we predict structural change upon point mutation?* International Conference on Intelligent Systems for Molecular Biology (ISMB 2011), Vienna, Austria, July 15-17 2011 (Talk).
- **Christian Schaefer**, Alice Meier, Burkhard Rost, Yana Bromberg. *Comprehensive nsSNP database with predictions and annotations of impact.* International Conference on Intelligent Systems for Molecular Biology (ISMB 2011), Vienna, Austria, July 15-17 2011 (Poster).

- **Christian Schaefer**, Avner Schlessinger, Burkhard Rost. *Protein secondary structure is robust under artificial evolution while protein disorder is not.* International Conference on Intelligent Systems for Molecular Biology (ISMB 2010), Boston, USA, July 11-13 2010 (Poster).

- **Christian Schaefer**, Avner Schlessinger, Burkhard Rost. *Protein secondary structure is robust under artificial evolution while protein disorder is not.* 7th annual Rocky Mountain Bioinformatics Conference 2009, Aspen, USA, December 10-12 2009 (Poster).

1 Introduction

Molecular evolution is considered as the driving motor for biological diversity. The interplay of mutation and selective pressure continuously adapts life to an ever-changing environment. The genetic makeup of each organism is the consequence of an ongoing sampling process over genetic configurations. While mutations occur by chance, selective pressure favors neutral or advantageous change over genotypes leading to deleterious phenotypes. It is this genetic variation that leads to different traits.

Most of the genetic variability in human is carried by single point mutations (Collins *et al.*, 1997, 1998). One of the greatest challenges in recent medicine and bioinformatics is to answer the question as to how slight genotypic variations could lead to different phenotypes. Of particular interest are traits pertaining to drug sensitivity or raised susceptibility to genetic diseases (Chakravarti, 1998; Fernald *et al.*, 2011).

1.1 Sequence determines structure

The genome of an organism contains the blueprint for one of its essential functional building blocks – proteins [1]. Their biosynthesis comprises a two-step enzymatic process and resembles a subsequent information flow from DNA (deoxyribonucleic acid) over mRNA (messenger ribonucleic acid) to protein (Crick, 1958, 1970). Protein-essential information is organized in coding parts of genes. The genetic code is universal to all organisms and combines three nucleotides to codons which unambiguously map to the 20 amino acids (e.g. Alberts *et al.*, 2002, chap. 6). The mapping is redundant in that more than one codon can code for one particular amino acid (e.g. proline is encoded by either one of the tri-nucleotides CCU, CCC, CCA, CCG on mRNA level). The consecutive order of nucleotides within coding regions defines the amino acid sequence of the polypeptide chain.

Amino acids consist of an amine-, a carboxyl-group and a variable side chain attached to the central carbon atom. Depending on the side chain, amino acids vary in their biochemical and physical properties, such as hydrophobicity, size and polarity. Through specific chemical interactions which involve side chains and

[1] Another group of functional macromolecules appears to be that of non-coding RNAs, their role is only now becoming understood (Mattick, 2009).

the peptide's main chain, residues close or distant in sequence come into spatial proximity. In this respect, a polypeptide chain could be seen as a sequence of these basic characteristics, since it is those that determine the intricate details in mutual interactions of residues but also the interplay with the solvent and bound ligands. The linear polymer folds into its unique three-dimensional structure (Doolittle, 1981; Zuckerkandl and Pauling, 1965), and the particular sequence of amino acids alone determines the structure of a protein (Anfinsen, 1973). It is this specific fold that enables a protein to fulfill its distinct functional role in the cell. In a nutshell, the primary sequence defines the structure which determines the protein's function.

The formation of secondary structure elements marks an essential prerequisite towards a stable unique three-dimensional structure for 'well-ordered' proteins. α-helices and β-sheets are the major examples for these structural motifs. Others are 3_{10}- and π-helices. They denote the basic macromolecular building blocks in proteins and are stabilized through hydrogen bonds (H-bonds) between carboxyl- and amine-groups of the backbone (Branden and Tooze, 1999). In consequence, the polarity contained within the backbone gets neutralized. In helices, H-bonds are formed between residues close in sequence, i.e. between the ith residue and residue $i + 3$, $i + 4$, $i + 5$ in case of the 3_{10}-, α- and π-helix, respectively. A β-sheet is stabilized through H-bonds between residues in extended parallel or anti-parallel strands that could be far away in sequence. The lack of any of these states is usually referred to as *coil*.

Essential to the folding process of proteins is the hydrophobic effect (Tanford, 1978). In soluble globular proteins, it is responsible for the segregation of non-polar residues into the protein interior and of polar residues to the surface (Guy, 1985). As a consequence, polar water molecules are largely excluded from the hydrophobic core while exposed side chains form favorable interactions with polar molecules in the solvent.

The overall tertiary structure is determined by the assembly of secondary structure elements. Their spatial arrangement is maintained mainly through medium and long range side chain contacts (rev. Chan and Dill, 1991; Dill *et al.*, 1995). The key role play interactions between side chains which stabilize the overall conformation partially through van der Waals forces between non-polar partners. Of further importance are H-bonds existing between two side chains or a side chain and the peptide backbone. Other stabilizing forces denote covalent disulphide bonds formed between two cysteine residues and electrostatic interactions of ionized groups. Especially the latter plays a dominant role for structure stability in thermophilic organisms (Korndörfer *et al.*, 1995; Perutz and Raidt, 1975; Taylor and Vaisman, 2010; Vetriani *et al.*, 1998; Yip *et al.*, 1995).

Proteins can form multimeric complexes which are referred to as quaternary

structure. The interactions between monomers are stabilized through the same atomic interactions as in secondary and tertiary structure. Complexes can occur in different manifestations (rev. Ozbabacan *et al.*, 2011). Homo-oligomers are composed of identical polypeptide chains while hetero-oligomers are composed of different ones. Obligate versus non-obligate interactions occur between monomers that are unstable or stable in their unbound states. Depending on their lifetime, complexes are classified into transient or permanent interactions.

Interactions between subunits occur through whole surface patches. These interfaces differ from each other in their hydrophobicity depending on the kind of interaction: Homodimers rarely function as monomers and their interaction surfaces are more hydrophobic than those of transient heterocomplexes which exhibit more hydrophilic properties (Janin and Chothia, 1990; Janin *et al.*, 1988; Jones and Thornton, 1996). In fact, amino acid properties of interfaces differ from those of the protein surface and furthermore appear to be even distinctive for each interface type (Ofran and Rost, 2003).

Another aspect of protein structure pertains to the interplay between well-ordered regions of regular secondary structure on one side and highly flexible and unstructured regions on the other. These parts have been termed as *intrinsically disordered* or *unstructured* regions. Protein disorder refers to segments in the polypeptide chain that do not adopt a well-defined three-dimensional structure in isolation. Residues in these regions rather exist in a variety of different conformations over time without exhibiting an equilibrium state. Differences in amino acid distributions between well-ordered and disordered regions imply that certain amino acids promote disorder (Dunker *et al.*, 2001). A significant feature of disordered regions is a depletion in hydrophobic residues and an abundance of parts that exhibit low sequence complexity (Romero *et al.*, 2001). Proteins can consist of disordered segments or can be disordered in total (Uversky *et al.*, 2005).

Protein disorder plays a dominant role in binding ligands and can occur as flexible domain linkers, in molten globule domains or in loopy protein ends and can be substantial to protein function (Dyson and Wright, 2005). They also participate in interfaces of protein complexes (Mészáros *et al.*, 2007).

The discovery of protein disorder amended the dogma that *sequence determines structure determines function.* The lack of a defined structure appears to be the key for a variety of biological processes that involve cell cycle control, gene regulation or signalling (Dunker *et al.*, 2002; Vucetic *et al.*, 2007). Its significance is only now becoming understood. In-silico predictions in whole genomes suggested that more than 30% of proteins in eukaryots contain regions that are devoid of well-ordered structure (Dunker *et al.*, 2000; Schlessinger *et al.*, 2011). Protein disorder may therefore contribute to the evolutionary means to transit from prokaryotic cells to more enhanced eukaryotic life.

1.2 Protein structure is conserved

Genes are exposed to evolutionary processes such as mutation under biological constraints. Proteins observable in contemporary organisms sustained purifying selection by retaining the individual's fitness and reproductive success (Kimura, 1983). In this regard, molecular evolution could be seen as a mutagenesis trial which generates new variants subjected to either fixation into population – or rejection.

Proteins evolve under functional constraints. Since protein function is established through protein structure, structure is also subjected to evolutionary selection. A particular condition crucial for the maintenance of globular proteins during the course of evolution is the upkeep of the closely packed interior of the hydrophobic core. This inflicts constraints upon the occurrence of acceptable amino acid substitutions within these regions: Change-in-bulkiness or hydrophobic-to-polar mutations are less expected to be observed than in exposed regions with more relaxed restraints (Franzosa and Xia, 2009). Consequentially, conservative amino acid exchanges that maintain relevant biochemical properties in constrained regions get enriched over time.

It has been long accepted that proteins sharing similar amino acid sequences fold into similar structures (Zuckerkandl and Pauling, 1965; Doolittle, 1981). To which extent could a sequence be changed until it looses its *wild-type* fold?

Relationships between sequence plasticity and structure maintenance have been studied by the example of the hemoglobins. The globin fold (Fig. 1.1) is present in virtually all three kingdoms of life and is encoded by orthologous genes. Indeed, phylogenetic analysis suggests lineage from an ancient ancestral gene (Vinogradov *et al.*, 2007). Sequence divergence ranges from one mismatch between human and gorilla (Goodman *et al.*, 1983) to 19% sequence identity between human and worm (Fig. 1.1A-C) to a level where homology could not be inferred from sequence. Despite this variety in sequence, globins fold into the same structure.

Early work investigated the variance in axial ratios and other coarse-grained structural features in crystals of hemoglobins (Reichert and Brown, 1909). Further advances in high-resolution X-Ray crystallography led to structures at atomic resolution and enabled detailed analyses of sequence variability in myoglobin and hemoglobin structures (Kendrew *et al.*, 1958; Perutz *et al.*, 1960). Through structural analyses in globins from a wide range of species (Perutz *et al.*, 1965; Lesk and Chothia, 1980), the following sequence-structure relationships became evident: (i) the structure's interior displays high amount of sequence variability while being restricted to non-polar residues, (ii) residues exposed to the surface are not subjected to those restrictions and exhibit changes from polar to non-polar residues and vice versa, (iii) mutations in buried residues are not constrained by size.

Analysis of these sequence constraints inflicted by the need of fold maintenance

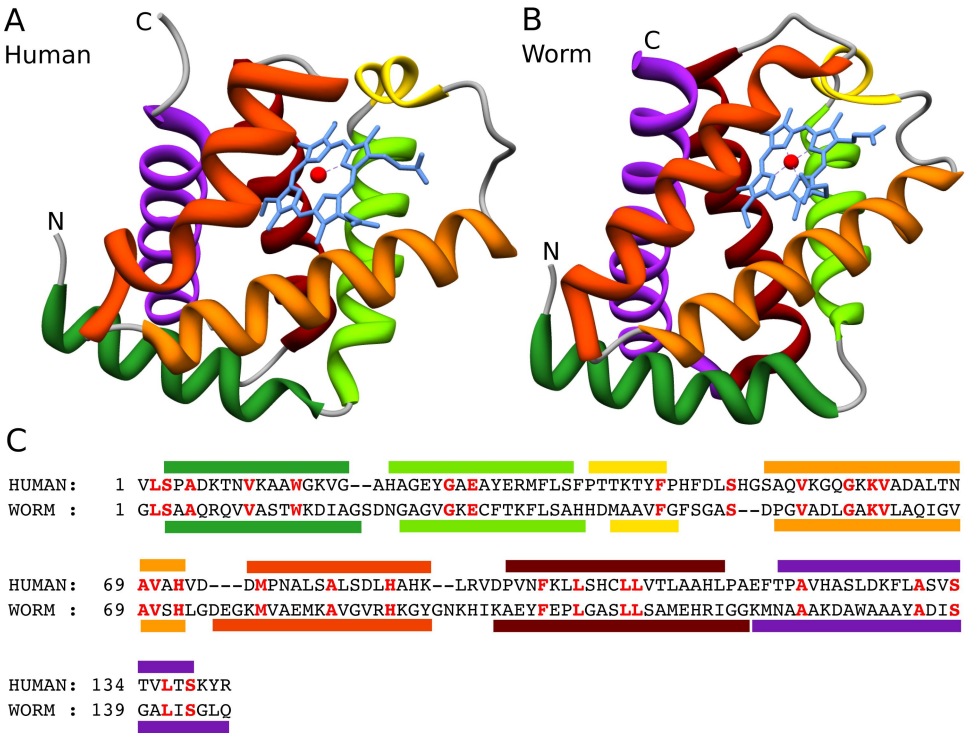

Figure 1.1: Globin fold conserved. Ribbon plots of two globin domains from (**A**) human (hemoglobin A, PDB 2W72A) and (**B**) marine worm (monomeric hemoglobin, PDB 1JL6A) and (**C**) sequence alignment; equivalent helices in both structures and their projection to sequence are represented in the same color; porphyrin-ring and attached iron colored in blue and as red sphere, respectively; identical residues in the alignment depicted in bold red. Despite a low sequence identity of 18.9%, both proteins assume the same structural configuration (main chain RMSD 1.63Å), i.e. orientation and connectivity between seven helices and maintenance of the heme binding pocket are conserved. Structures rendered with Chimera (Pettersen *et al.*, 2004), alignment calculated with FATCAT (Ye and Godzik, 2003).

indicate that globins react on mutations in the hydrophobic core by rigid-body shifts (Lesk and Chothia, 1980). These movements are dissipated to hinge regions connecting helices while retaining the heme binding site which is crucial for globin function. A lot of interacting side-chains involved in helix-helix interfaces appear to have co-mutated, thus retaining the densely-packed environment essential for structural integrity (Lesk and Chothia, 1980). Indeed, shifts in backbone (Alber *et al.*, 1988; Weaver *et al.*, 1989) and side-chain movement (Eyal *et al.*, 2003; Ponder and Richards, 1987) seem to provide means to enable proteins to accommodate for mutations while retaining their fold.

Surprisingly, structural restraints in partially and fully buried sites seem to be even more relaxed than anticipated. Recent work indicates that up to four hydrophobic-to-charged substitutions in mutual structural vicinity could be tolerated, provided that the backbone of the wild-type protein exhibited a certain folding stability (Isom *et al.*, 2010; Garcia-Seisdedos *et al.*, 2012).

Do these observations merely fit to a select group of protein families or is the high structural tolerance for sequential change the rule? What are the restraints on sequence similarity? The increase in known structures at atomic resolution in the Protein Data Bank (PDB, Berman *et al.*, 2000) led to a gain in coverage of distinct folds. Through this, it became possible to get deeper insight into general aspects of the mapping between sequence and structure space by large-scale pairwise structural comparisons.

Sequence identity between two naturally evolved proteins can go down to 35% to infer a similar structure (Rost, 1999; Sander and Schneider, 1991). One exception to this rule was found recently (Roessler *et al.*, 2008).

The vast majority of similar structure pairs, however, appears to have on average only ~10% of their sequence in common (Rost, 1997). Whether this observation could be solely attributed to either convergent or divergent evolution remains to be under speculation. A combination of both effects appears to be likely (Rost, 1997).

Apparently, protein sequences contain an intrinsic redundancy suggesting that a protein fold does not strictly depend on a distinct amino acid sequence (Baase *et al.*, 1992). It is rather the pattern of biochemical properties conserved within a protein that contributes to the overall structure and that needs to be maintained throughout evolution (Bowie *et al.*, 1990, 1991; Kamtekar *et al.*, 1993; Shakhnovich and Gutin, 1991). Substitutions that support the intricate atomic interactions crucial for maintaining structure and function are tolerable. Evolution had enough time to remove severely deleterious mutations from the pool of contemporary sequences. In this respect, the vast majority of mutations observable today are those that posed least damage to protein structure.

In short, protein structure is more conserved than protein sequence.

1.3 Mutations in structural hotspots lead to deleterious effects

Direct observation of structural effects due to amino acid exchanges is cumbersome since each mutant protein must undergo the time consuming and challenging crystallization process. The complexity further increases when a set of candidate mutations are to be screened independently for their structural impacts. Despite

these obstacles, the PDB (Berman *et al.*, 2000) contains a few mutant structures (Eyal *et al.*, 2001) and structural effects occurring upon changing a single residue have been observed. On the other hand, measuring changes in protein stability provides an alternative to ascertain effects due to amino acid exchanges.

Examples of effects include polymerization of serine protease inhibitors (rev. Gooptu and Lomas, 2009), the formation amyloid fibrils in p53 (Galea *et al.*, 2005), switching from mainly-beta to an all-alpha fold (Alexander *et al.*, 2009), or affecting the structural stability in general (Shirley *et al.*, 1992; Betz, 1993).

Single point mutations that lead to a severe change in biophysical properties at the mutated site could severely affect structure and stability. Important examples are small-to-bulky mutations that could even have a influence beyond the local structural neighborhood of the changed site (Alber and Matthews, 1987; Buckle *et al.*, 1996; Dao-pin *et al.*, 1991; Eriksson *et al.*, 1992; Liu *et al.*, 2000; Sandberg and Terwilliger, 1989; Xu *et al.*, 1998) as well as changes of hydrophobicity in solvent inaccessible sites (Matthews, 1993; Shortle *et al.*, 1990).

Local structure in the vicinity of glycines could be especially susceptible to mutations. With a single hydrogen atom as side chain, glycine is able to sample a much larger space in backbone dihedral angles than other amino acids (Creighton, 1993), a feature that makes it abundant in reverse turns (Rose *et al.*, 1985). These sites are prone to certain mutations against bulky and less flexible residues (Pakula *et al.*, 1986).

Mutations that involve proline denote another class that could lead to significant structural effects. Proline takes on a special position amongst the 20 amino acids in that the terminal end of its side chain forms a cyclic structure through a covalent bond with the amine-group. This has two implications. First, its ϕ dihedral angle is nearly rigid and is constrained to values around -60°. Second, the lack of the hydrogen at the amine group results in the loss of an H-bond donor. These effects are considered as reasons for the depletion of proline in α-helices (Chou and Fasman, 1978; Richardson and Richardson, 1988; Schimmel and Flory, 1968). Structural impacts in α-helices induced by single amino acid exchanges that introduced proline have been observed (Gray *et al.*, 1996; Hecht *et al.*, 1983; Shortle and Lin, 1985). The reverse effect has been studied in T4 lysozyme where the exchange of a wild type proline led to the extension of an α-helix (Alber *et al.*, 1988).

Since protein function is established through structure, function should not be unaffected by certain conformational rearrangements. A severely changed or disrupted function may furthermore lead to a disease phenotype. Characteristics of function-critical sites are manifold and depend on the particular role a protein plays in the cell. Sites that realize catalytic activity consist of only a small number of residues while the binding of large molecules such as DNA or other proteins

involves whole surface patches. Accordingly, a small structural effect may be sufficient to reduce the binding affinity and in consequence disrupt function, while in other cases larger structural rearrangements may be necessary.

Structural and functional impacts induced by disease-related amino acid exchanges have been studied in detail. For example, mutations in four active site residues of glucose-6-phosphatase inactivate its enzymatic activity and cause glycogen storage disease (Lei *et al.*, 1993; Shieh *et al.*, 2002). Several conformational changes have been observed in DNA-binding regions of the tumor suppressor p53 which were induced by cancer-related mutations (Joerger *et al.*, 2006). The replacement of a buried hydrophobic methionine by a positively charged arginine results in a complete destabilization of a decarboxylase which leads to a skin-related disease (porphyria cutanea tarda, Mendez *et al.* (1998)). Two disease-related mutations in the copper-binding region in a superoxide dismutase result in a significant activity decrease and promote the onset of familial amyotrophic lateral sclerosis (Ferraroni *et al.*, 1999; Ratovitski *et al.*, 1999).

Comparisons with benign variants could provide insights into the significance of structural regions and features that play a dominant role in disease development. However, solved crystal structures of mutant proteins with an associated disease are rare. Thus the protein diversity is low and conclusions drawn may not be significant. Alternative protocols usually involve the mapping of mutant residues onto known structures through a simple alignment between mutant sequence and sequences of candidate structures. More sophisticated methods conduct an automated homology modeling of the wild type protein and perform a refinement of the structural neighborhood after introducing the mutant side chain. These approaches allow for a broader coverage however at the cost of structural details and accuracy.

Despite these deficiencies, interesting insights on coarse-grained level could be gained. Investigations suggested that disease-related mutations occur more often at solvent inaccessible positions (Sunyaev *et al.*, 2000; Gong and Blundell, 2010; Wang and Moult, 2001). Furthermore, they were shown to be enriched in β-sheets (Ferrer-Costa *et al.*, 2002; Gong and Blundell, 2010) and could lead to an overpacking of the protein core or result more often in an H-bond loss (Gong and Blundell, 2010; Wang and Moult, 2001). These rules may offer valuable means for predicting the disease-relatedness of new discovered yet unannotated variants.

1.4 Effects of most variants unknown

During the past years, new high-throughput sequencing methods have led to a literal explosion of genomic data while significantly lowering the costs (rev. Kircher and Kelso, 2010). International collaborative endeavors such as HapMap (Inter-

national HapMap Consortium, 2003), the 1000 Genomes Project (1000 Genomes Project Consortium, 2010) and dbSNP (Sherry *et al.*, 2001) collect data on human genetic variants in large scale.

The vast majority (~90%) of human sequence variants are single nucleotide polymorphisms (SNPs) (Collins *et al.*, 1998). Of particular interest are nsSNPs (non-synonymous SNPs) since they alter an amino acid in the gene product and could affect protein structure and function (Sunyaev *et al.*, 2000) with further consequences for disease (s. section 1.3).

However, phenotypic effects for the vast majority of human nsSNPs are not known. Functional implications have been determined for only ~3% and ~4% of all known nsSNPs are associated with a disease (Fig. 1.2). Genetic association studies strive for finding causative variants by detecting significant genetic differences between groups of disease affected and healthy individuals. However, broad genetic variability within a group could complicate the delineation of disease-related variants. Furthermore, genetic disorders are often complex in that they are caused by variations in more than one gene, each with a small individual effect. Environmental factors may play an additional role Collins *et al.* (1997, 1998).

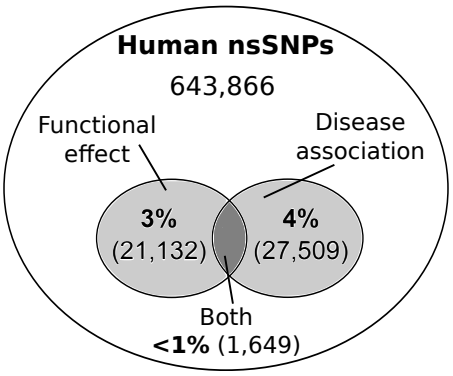

Figure 1.2: Majority of human variants without annotation. Only a small fraction of human nsSNPs has information on their functional impact (3%); 4% is known to be disease-related and <1% have a functional and disease annotation. (Figure adapted from Schaefer *et al.* (2012b).)

Hence, there is need for in-silico methods that determine the likelihood of effect at high accuracy. Evolution provides the means to ascertain whether a mutation is likely to be deleterious or not. On the one hand, variants leading to a severe or lethal effect are unlikely to be passed to the offspring generation and therefore are rarely observed. On the other hand, mutations in less constrained regions are evolutionary more accepted since they pose little selective disadvantage and therefore accumulate in the population. Evolutionary profiles obtained from sequence align-

ments provide means to gauge sites that developed under structural or functional constraints in homologous proteins.

Early in-silco methods were based solely on sequence profiles and predicted whether a mutation is deleterious or not based on the mutant frequency at the respective position of the profile (Ng and Henikoff, 2003; Sunyaev *et al.*, 2001). The inclusion of structural information or the use of machine learning algorithms led to more sophisticated prediction methods (Bromberg and Rost, 2007; Bao and Cui, 2005; Capriotti *et al.*, 2006, 2005; Chasman and Adams, 2001; Krishnan and Westhead, 2003; Saunders and Baker, 2002; Wang and Moult, 2001). Relevant properties often include changes in biophysical properties such as hydrophobicity or bulkiness, structural features such as secondary structure and disorder propensities or solvent accessibility.

Different methods operate on different perceptions of phenotypic effect, such as change in structure (Wang and Moult, 2001), stability (Capriotti *et al.*, 2005), function (Bromberg and Rost, 2007; Bao and Cui, 2005; Chasman and Adams, 2001) or pathogenicity (Capriotti *et al.*, 2006). All have their weaknesses and strengths when applied to different sets of mutations and their predictions on the deleteriousness of a mutation might not be conclusive. Nonetheless, in-silico methods trained on experimental data offer a first step in filling the annotation gap of nsSNPs and provide means to prioritize further experimental investigation. The combination of predictions might offer further insights into the molecular details of disease mutations.

1.5 Thesis incentives and outline

The work at hand is structured as follows, sections and respective underlying publications are highlighted.

The main objectives of this dissertation were two-fold. First, we developed a machine-learning model aimed at predicting impacts on local protein structure induced by a single amino acid exchange. Second, we showed that predicted effects on structural and functional level were enriched in disease-associated mutations. Based on these findings, we tested whether combined predictions of both effects could lead to an accurate assessment on whether a point mutation induces disease or not.

In a preliminary analysis, we mutated protein sequences to random-like strings of amino acids and investigated changes in predicted secondary structure and predicted protein disorder (sections 2.1, 3.1, 3.2). Our analysis revealed a surprisingly high robustness in content and length of secondary structure elements. We also observed significant structural changes pertaining to switches between helices and strands and the dis- and reappearance of disordered regions (Schaefer *et al.*, 2010).

Furthermore, we introduced a new concept of structural change that allowed for the compilation of a large training set consisting of structural neutral and effect cases. With these data at hand, we successfully induced a logistic regression-based method to distinguish between structural neutral and non-neutral mutations solely on sequence-based information (sections 2.2, 3.3, Schaefer and Rost (2012)).

In a comprehensive collection of disease-associated nsSNPs (section 2.3, Schaefer *et al.* (2012b)), we predicted effects on structure and function. We observed that strongly predicted effects on either structure or function were enriched in disease-related variants. This led to the indication that severe impacts on molecular level raise the likelihood for a mutation to be deleterious (sections 3.4, 3.5, Schaefer and Rost (2012); Schaefer *et al.* (2012a)).

Motivated by these results, we tested both methods individually and in concert with respect to their ability to distinguish between non-deleterious and disease-related mutations. We found that predicted functional effect alone sufficed to accurately predict a mutation to be deleterious or not (section 3.6, *unpublished*).

2 Materials and Methods

2.1 Monitoring structural change under in-silico mutation

Sequences from globular and disordered proteins

We studied the persistence of predicted secondary structure and predicted disordered regions under sequence changes. For this purpose, we used two protein sequence databases as the basis for our analyses. The first set comprised sequences of well-ordered globular proteins extracted from the Protein Data Bank (PDB, Berman *et al.*, 2000). We only considered proteins solved by X-Ray diffraction. The second set consisted of sequences containing disordered regions taken from the DisProt database (Vucetic *et al.*, 2005) (version 4.9).

In general, sequence databases contain several kinds of bias. One can be attributed to the over-representation of certain protein families. To ensure results that were unbiased towards any redundancy that may have been inherent in our initial sequence collections, we applied UniqueProt (Mika and Rost, 2003) and filtered both sets at a homology threshold of an hssp value >10 (Rost, 1999; Sander and Schneider, 1991). This specific value was chosen empirically and based on experience drawn from earlier investigations. It roughly corresponds to 30% pairwise sequence identity for alignments longer than 250 residues. After this procedure, the redundancy-reduced sets consisted of 1,369 protein sequences from the PDB and 374 from DisProt.

Mutation protocol

We mutated each sequence in both redundancy-reduced sets step by step into random-like strings of amino acids. At each new step, we arbitrarily picked 10% of amino acids in the sequence from the previous step. Then, we mutated each chosen amino acid X to amino acid Y with a particular substitution probability p_{XY}. We calculated p_{XY} adhering to two alternative schemes. First, we used the substitution matrix PAM120 (Dayhoff *et al.*, 1978) and mutated according to PAM120 probabilities. Second, we interpreted the background amino acid distribution of the respective underlying sequence database as substitution probabilities. For each

native sequence and each substitution scheme, we created a 'mutation trajectory' consisting of overall 69 of these single steps.

In addition, we used each of both redundancy-reduced sets as the basis for creating a corresponding set of random amino acid sequences. We retained the following characteristics of the original sets: (i) the amino acid composition, (ii) the distribution of sequence lengths and (iii) the amount of sequences. The purpose of these random sequences was to test whether we reached a state of convergence during our in-silico mutation protocol, that is, to ensure that we eradicated any 'memory' contained within the native sequences.

Further sequence data used

Proteins in the PDB may not resemble the full universe of observable protein families. One reason is the difficulty that lies in the crystallization process of membrane-bound proteins (Carpenter *et al.*, 2008) which makes them strongly underrepresented in the PDB. Furthermore, some protein families may not be represented at all. To overcome this problem and to reduce further bias, we used an additional set of 33,812 protein sequences taken from RefSeq (Pruitt *et al.*, 2005). These sequences represented the entire human proteome.

We compared results from our in-silico mutation procedure to naturally evolved homologous proteins taken from the HSSP database (*Homology-derived Secondary Structure of Proteins*, Sander and Schneider, 1991). For each of the 1,369 sequences in the redundancy reduced set of globular proteins, we randomly extracted ten of their homologs from HSSP such that we evenly covered the whole range of available sequence diversity. In addition, we monitored the sequence identity for each homologous pair.

Determining secondary structure and disordered regions

For each native PDB sequence and every mutated sequence along the mutation trajectory, we predicted the secondary structure content through PROFsec (Rost, 2005). We run predictions in 'sequence-mode', that is, we did not compile sequence profiles as input but rather presented the raw sequence to the method. We were explicitly interested observing effects based on small sequence changes, which would have been obscured by the use of profiles otherwise. This procedure resulted in a reduced accuracy of ~68% Q3 (percentage of correctly predicted residues in one of the three states helix, strand, other) compared to the mode using evolutionary profiles (~72% Q3) (Rost, 1996, 2005).

In addition to predictions, we determined the observed secondary structure in native PDB sequences as calculated by DSSP (Kabsch and Sander, 1983). We converted the initial eight DSSP states into three, representing helix, strand and

other (Rost and Sander, 1993; Rost, 1996; Andersen *et al.*, 2002). In each mutation step (i.e. after changing 10% of the sequence), we monitored the sequence identity compared to the native sequence, the relative content of residues predicted in the states helix and strand, and the segment length of predicted helices and strands.

For each native and mutated sequence derived from the DisProt set, we predicted short and long disordered regions through IUPred (Dosztanyi *et al.*, 2005a,b). The method accepted as input the raw protein sequence and provided three modes optimized to predict either long or short disordered regions or residues in globular domains. We chose the former two options to predict long and short disorder, respectively. For each mutation step, we monitored sequence identity compared to the native sequence, the relative content of residues predicted in short and long disorder, and the segment length of these regions.

2.2 Learning the structural effect upon a single residue exchange

2.2.1 Pairs of pentamers

One major incentive of this thesis was the development of a machine-learning based method to predict the local structural effect that occurs upon the exchange of a single amino acid. To this end, we created a training set consisting of pairs of pentapeptides by adhering to the following protocol.

Based on 146,296 protein chains taken from the PDB (Berman *et al.*, 2000) (July 2010), we created two separate sequence sets, both redundancy-reduced to different levels of sequence identity. The first set (referred to as 'cdhit98') resulted from a sequence clustering at 98% identity threshold using CD-HIT (Li and Godzik, 2006) and reduced the original set to 24,890 sequences. The second set ('hval0') consisted of 3,767 chains resulting from filtering at an hssp value >0 (Mika and Rost, 2003; Rost, 1999; Sander and Schneider, 1991). This level of redundancy corresponds to ~20% maximal pairwise sequence identity for an alignment length of over 250 residues.

We sampled overlapping fragments of five consecutive residues (pentamers) from each protein chain in both sets and paired each pentamer from the first set ('cdhit98') with each of the second set ('hval0'). We considered pairs that (i) only contained standard residues, (ii) had no gaps in their backbone (i.e. chain breaks with peptide bond length >2.5Å (Kabsch and Sander, 1983)) in either pentamer, (iii) contained no alternative sets of atomic coordinates, (iv) originated from proteins with over 30% pairwise sequence identity, and (v) differed only in their central amino acid. This procedure resulted in 35,533 pairs of pentamers.

Our objective was to evaluate a possible conformational change caused by the

central mismatch residue. We did not know beforehand about the most effective way to capture a structural effect nor its extent. A commonly used metric to ascertain the similarity of two protein structures is the root mean square displacement (RMSD) determined after optimal superposition of both structures (Kabsch, 1976, 1978). We calculated the backbone RMSD over all C_α atoms (McLachlan algorithm (McLachlan, 1982), as implemented in the program ProFit [1]).

The range of observed RMSD values started at values close to zero (structural very similar) and was not bound by an upper limit (large values translate to structural very dissimilar). We mapped large values to a positive class (structural effect) and small values to a negative class (structural neutral). The RMSD thresholds were chosen such that we obtained even amounts between effect and neutral pairs. The specific cutoffs were <0.2Å for the negative and >0.4Å for the positive class. These thresholds assigned 13,675 pentamer pairs to the negative class and 12,046 to the positive class. All pairs in between this range were discarded. For each neutral and effect pair we randomly designated one fragment as 'wild type' fragment and the central mismatch residue of the other fragment as the mutant amino acid.

2.2.2 Constructing a ground set of features

We did not know beforehand which features were significant for the task at hand. Therefore, we adhered to common practice in the field and created an excessive baseline set of potential features which we subjected to a forward selection procedure afterwards (s. section 2.2.4). Based on knowledge gained during the development of a similar method (Bromberg and Rost, 2007), we extracted all sequence-based features from PredictProtein (Rost *et al.*, 2004). This method constitutes a wrapper that combines several independent methods for predicting structural and functional features, each operating on protein sequences. During our feature constructing process, we distinguished between three conceptually different classes of features: *global features* describing the global characteristics of the protein sequence in its whole, *local features* describing one particular pentamer and its immediate sequence neighborhood, and *difference features* that explicitly described sequence-derived aspects by which wild type and mutant amino acid differed.

Global features

The sequence length was encoded by four distinct values, each representing a certain length interval (1-60, 61-120, 121-180, >180 consecutive residues). The specific value that represented the length was set to 0.5, values below and above were set to 1 and 0, respectively. We represented the relative frequency of standard

[1] http://www.bioinf.org.uk/software/profit/

amino acids by 20 values. Three further values encoded the relative content in predicted helix, strand and loop states and additional three values encoded the relative content in predicted buried, intermediate and exposed residues (Rost, 1996, 2005).

Local features

We tested several sequence windows (1, 5, 9, 13, 17, 21) of consecutive residues in the wild type fragment, each centered around the position of change exchange. All feature values were normalized to fit the interval [0,1].

Basic features we considered were six different biochemical and structural properties of standard amino acids: mass, volume (Zamyatnin, 1972), hydrophobicity (Kyte and Doolittle, 1982), C-beta branching (Betts and Russell, 2003), helix breaker (only proline) and electric charge of side chain.

We extracted evolutionary profiles from PSI-BLAST (Altschul *et al.*, 1997) runs (options: -j 3 –b 3000 –e 1 –h 1e-3) against a redundancy-reduced sequence database consisting of UniProt (Bairoch *et al.*, 2005) and PDB (Berman *et al.*, 2000). Of interest were the position specific scoring matrices (PSSMs), relative amino acid frequencies and the information content per alignment position. As an alternative to PSSM, we also applied the PSIC method (position-specific independent counts, Sunyaev *et al.*, 1999), which has been already used elsewhere with success (Bromberg and Rost, 2007; Sunyaev *et al.*, 2001).

We considered the following predicted structural and functional features to be important for our setting: secondary structure (Rost and Sander, 1993, 1994) and solvent accessibility (Rost and Sander, 1994; Rost, 1996, 2005), protein flexibility (Schlessinger *et al.*, 2006), protein disorder predicted through IUPred (Dosztanyi *et al.*, 2005a) and MD (Schlessinger *et al.*, 2007b,a, 2009), protein-protein interaction sites (Ofran and Rost, 2003, 2007b,a) and DNA-binding residues (Ofran *et al.*, 2007).

The majority of these methods returned a discrete prediction, denoting the particular state of a residue (e.g. disordered or not). We represented two-state predictions (disorder, protein and DNA interaction) and three-states predictions (secondary structure states helix, strand, other and solvent accessibility states buried, intermediate, exposed) through combinations of 1 (presence of a state) and 0 (absence of a state). In addition, we augmented these discrete predictions with their raw scores, reflecting the strength and reliability of the prediction. Protein flexibility was predicted as a numerical value only.

Furthermore, we incorporated information about the position of change relative to a protein domain in our feature set. For instance, a hydrophobic-to-polar exchange may lead to a significant structural effect through rearrangements in the structural neighborhood when occurring within the hydrophobic core of a domain.

Whereas the effect in a flexible surface loop may be less pronounced. We used four different pieces of information derived from sequence alignments against the Pfam-A database (Finn *et al.*, 2010) and produced by HMMER3 (Finn *et al.*, 2011): the information about whether the residue change occurred inside a domain, the evolutionary conservation of that position within the domain alignment, how well the residue fitted into the alignment position and the posterior probability of that match.

Difference features

We anticipated that a structural difference in a pair of pentamers may be induced by the underlying characteristics of the differing central amino acids. Hence, we incorporated several such properties into our baseline feature set. A particular feature difference was encoded through its absolute value and sign, reflecting strength and direction of change.

The following properties were encoded in that respect: Change in any of the six biochemical amino acid propensities, difference in conservation scores (PSSM, relative frequency, PSIC), change in IUPred predictions for both short and long disorder, change in predicted secondary structure and solvent accessibility. For the latter two we ran PROFphd on raw sequence rather than sequence profile. Although this mode resulted in reduced prediction performance, it allowed us to observe an actual difference in the prediction outcome, which would have been disguised by the use of sequence alignments otherwise (s. also section 2.1).

2.2.3 Machine-learning algorithm

We chose to apply a logistic regression based approach to our problem. Logistic regression is a parameter-free method that could lead to similar predictive power as support vector machines while being significantly faster during model building and testing (Fan *et al.*, 2008). We adhered to an implementation realized within the LIBLINEAR package (L2-regularized logistic regression, dual) (Fan *et al.*, 2008).

2.2.4 Selecting most predictive features

Irrelevant features often lead to raised computational cost and could even deteriorate predictive performance of the classifier (Guyon and Elisseeff, 2003). Therefore, we concentrated the testing and training of our classifier on the most significant features.

A straightforward procedure is the forward feature selection in a wrapper approach. Here, in an iterative process the feature set is gradually built up by adding single features that maximally raise the predictive performance of an underlying

machine-learning model (Guyon and Elisseeff, 2003). However, it is imperative to conduct this selection and subsequent assessment of the classifier's generalization ability on two distinct datasets. The objective is to prevent an overly optimistic performance estimation (Smialowski *et al.*, 2010). In addition, we took further precautions and ensured that no sequence homology existed between any subset used during feature selection and performance assessment (s. below).

We separated one fifth from set of pentamer pairs (s. section 2.2.1) by maintaining an even distribution between structural neutral and non-neutral pairs. Furthermore, we ensured that the pairs were derived from sequences without significant sequence homology (based on an e-value $> 10^{-3}$) to sequences in the remaining four fifth of pairs. The resulting 5,125 instances comprised 2,243 structural effect and 2,882 neutral pairs. These were further separated into ten subsets; class distribution and sequence dissimilarity (e-value $> 10^{-3}$) between all ten sets were maintained. We used nine such sets for training a logistic regression model and tested its performance on the remainder. We rotated ten times over all sets such that each instance served once during testing and training.

Before each new rotation, a set of features for training and testing the model was determined by the following iterative protocol. We started with one feature and established its predictive performance during one complete rotation as explained above. We did that for all global and difference features as well as every combination between local features and window lengths. We measured feature performance by means of average AUC (area under the receiver-operator curve) derived from rotating ten times over the testing folds. The best performing feature was automatically included for the subsequent evaluation of the remaining features. We stopped this forward selection after no further increase in average AUC>0.001 was observed.

2.2.5 Estimating predictive performance

The remaining four fifth of pentamers that had not been used during feature selection (overall 20,596) were divided into ten subsets respecting the same conditions as explained above. We conducted a 10-fold cross validation similar to that during feature selection. Using the most predictive features, we trained a logistic regression model on nine tenth of data and tested its performance on the remaining one tenth. We rotated ten times such that each instance served once during testing and training. After each round of testing, we monitored the following performance measures.

We used TP (true positives) to denote pairs that were correctly predicted to change structure (positive) and FP (false positives) to refer to neutral pairs wrongly predicted as change. Similarly, TN (true negatives) described the correctly predicted neutral pairs (i.e. no structural change) and FN (false negatives)

were structure-changing pairs incorrectly predicted as being neutral.

Our logistic regression model yielded a probability p for an instance to be structurally non-neutral rather than a discrete class label. In consequence, the particular values for the four measures depended on the specific probability threshold chosen to define the decision boundary between the two classes (i.e. effect versus neutral). By iterating over the whole range of possible cutoffs, we obtained a ROC-curve (Receiver Operating Characteristic) determined by pairs of *True Positive Rate* (TPR) and corresponding *False Positive Rate* (FPR). In a similar way, we created a ROC-like space of accuracy-coverage pairs for each of the two classes.

These measures are defined as follows:

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

The two measures AUC (area under the ROC-curve) and $Q2$ (two-state accuracy), both averaged over ten rounds of training and testing, served as single performance estimators.

$$Q_2 = \frac{TP + TN}{TP + FP + TN + FN}$$

Finally, we monitored class-specific values for $Accuracy_{Effect}$, i.e. the accuracy for the class 'structural change', $Accuracy_{Neutral}$ (accuracy for the class 'neutral'), $Coverage_{Effect}$ (coverage for class 'change') and $Coverage_{Neutral}$ (coverage neutral) defined by:

$$Accuracy_{Effect} = \frac{TP}{TP + FP}$$
$$Coverage_{Effect} = \frac{TP}{TP + FN}$$
$$Accuracy_{Neutral} = \frac{TN}{TN + FN}$$
$$Coverage_{Neutral} = \frac{TN}{TN + FP}$$

2.3 Collecting annotated single amino acid exchanges

A variety of public databases exist that contain information about genetic variants and their mappings onto mRNA or protein sequences. However, the sole knowledge

of genetic variation without effect annotation is hardly useful. Some attempts exist to unify and combine interesting annotation (Bairoch *et al.*, 2005; Kawabata *et al.*, 1999; Sherry *et al.*, 2001). Nonetheless, significant pieces of information still remain scattered across the universe of available mutant collections. Another problem relates to the mapping of point mutations to different protein identifiers although they reference the same sequence.

We developed a database based on a comprehensive table schema with the incentive to store and update relevant information on nsSNPs, their observed and predicted (Bromberg and Rost, 2007; Ng and Henikoff, 2001) effects as well as possible disease consequences. We collected data on protein sequences, single amino acid exchanges and associated consequences from four major sources of genetic variation:

(i) Swiss-Prot (Boeckmann *et al.*, 2003) denotes the central database for protein sequences and a variety of manual and reviewed annotations. Swiss-Prot is augmented by SwissVar (Yip *et al.*, 2008), both providing in their entirety information on natural variants and artificially created mutants, annotated with functional effects, the structural environment as well as disease consequences.

(ii) *The Protein Mutant Database* (PMD, Kawabata *et al.*, 1999) stores functional annotation on amino acid exchanges extracted from scientific publications.

(iii) The largest collection of several kinds of genetic variation (such as indels, copy number variations, SNPs) constitutes the *Single Nucleotide Polymorphism database* (dbSNP, Sherry *et al.*, 2001). Disease associated variants - in particular nsSNPs - contain references to OMIM (*Online Mendelian Inheritance in Man*, Amberger *et al.*, 2009), a knowledge base of human genetic diseases.

(iv) A comprehensive collection of human genetic variation is provided by the *1000Genomes Project* (1KG, 1000 Genomes Project Consortium, 2010). Although 1KG does not provide functional or disease annotations, we were interested in the frequencies with which nsSNPs occur in the human population. We mapped genomic nsSNPs onto protein sequences obtained from RefSeq (Pruitt *et al.*, 2005) through ANNOVAR (Wang *et al.*, 2010).

Irrespective of the original database, we treated two protein sequences as identical if the only difference was either a single residue exchange anywhere in sequence or a missing methionine residue at the beginning of either sequence. We represented each wild type sequence as its md5 checksum (as described e.g. in Smith *et al.*, 2005). This allowed us to unambiguously and efficiently correlate mutations originating from different sources but referencing the same canonical sequence.

For each point mutation, we assigned functional effect annotations (taken from Swiss-Prot, SwissVar, PMD) and disease consequences (SwissVar, PMD, OMIM), if available. Overall, we collected 1,362,793 unique single amino acid exchanges in 158,004 protein sequences coming from 2,684 organisms. The top five contrib-

utors were human, mouse, rice, cow and rat. Human nsSNPs accounted for 47% (643,866) of all mutants, out of which 3-4% were either functionally annotated (21,132) or had an associated disease (27,509). Less than 1% (1,649) contained information on both.

2.3.1 Method evaluation data

Based on this comprehensive collection of annotated variants and further external data, we compiled different sets:

Effect on function. The first set denoted the training set of SNAP (Bromberg and Rost, 2007). It consisted of a collection of point mutants out of which 39,397 were annotated as having an effect on protein function and 40,756 were annotated as functionally neutral. These mutations have been observed in 6,133 proteins.

Effect on stability. The second set comprised 1,297 single amino acid exchanges in 47 proteins (Capriotti *et al.*, 2005; Kumar *et al.*, 2006). Each mutant was experimentally annotated with respect to its effect on protein stability expressed as a change in Gibbs free energy $\Delta\Delta G$. Overall 647 mutations were considered as changing stability significantly (non-neutral) that led to a $\Delta\Delta G <$ -1 kcal/mol (stabilizing effect) or >1 kcal/mol (destabilizing effect). The remaining 650 mutations exhibited values within the $\pm$ 1 kcal/mol range and were assigned to the neutral class, that is, they did not change stability significantly.

Effect on disease. The third set consisted of 26,367 disease-annotated and 40,756 non-deleterious variants observed in overall 7,486 proteins.

2.3.2 Disease-related and functional-effect mutations

We created five different subsets of nsSNPs: (i) Set of *disease-related + observed effect* mutations: We collected 1,105 nsSNPs (from 217 proteins) that were annotated to be both disease-causing and functionally non-neutral. (ii) Set of *disease-related* mutations: We obtained 26,404 nsSNPs in 3,419 proteins that had an disease-association but no functional effect. (iii) Set of *observed effect* mutations: We collected 36,317 mutants in 3,790 proteins with functional effect but no disease association. (iv) Set of mutations with *unknown disease relation*: We extracted 251,414 variants in 28,913 proteins without known disease associations. (v) Set of *random* mutations: We randomly selected one mutation in each of the 28,913 proteins from the set of mutants of unknown disease relation such that the mutated

position was maximally distant from any other mutation observed in the given protein.

2.4 Predicting functional effects in disease-related mutations

We studied correlations between functional effects and disease. For that purpose, we used SNAP (Bromberg and Rost, 2007) and predicted the functional impact in five different data sets of nsSNPs (s. section 2.3.2). SNAP provides binary classifications (functional effect versus neutral) and a raw prediction score that offers a more elaborate view on the prediction outcome. Scores range from -100 (strongly predicted as neutral) to 100 (strongly predicted to change function); the distance from the binary decision boundary (0) measures the reliability of the effect. Essentially, stronger predictions are also more reliable, i.e. the higher the score, the more likely the mutation impacts function (Bromberg and Rost, 2007, 2008; Bromberg *et al.*, 2009). For a small data set, SNAP scores were shown to correlate with the severity of change; i.e. high (positive) SNAP scores relate to more severe functional effects (Bromberg and Rost, 2007, 2008; Bromberg *et al.*, 2009).

For many prediction methods developed in our group (protein-protein binding (Ofran and Rost, 2003, 2007b,a), protein-DNA binding (Ofran *et al.*, 2007), backbone flexibility (Schlessinger *et al.*, 2006)), the strength of an effect correlated with prediction strength, e.g. ISIS predicted binding hot spots stronger than other residues involved in the interaction (Ofran and Rost, 2007b). Although we never used the strength of an effect to train our methods, this correlation appears intuitive: stronger effects are more consistent and therefore become stronger carved into the machine-learning model.

2.5 Depicting results through box plots

Throughout our data analyses, we used box plots (Tukey, 1977; McGill *et al.*, 1978) for a comprehensive representation of results. Box plots provide means to condense interesting pieces of information in a distribution of observations to a single graphical element. This makes it possible to compare the basic statistical behavior of multiple sets of data points originating from different experiments. Box plots extend the common representation of a distribution through an average value and its spread, usually given by the standard error around the mean.

Here, a distribution is depicted by its basic characteristics, that is, its first three quartiles that build up the box. The lower and upper edges of the box depict the

first and third quartile of the data, respectively. The length of a box is referred to as the interquartile range of the distribution. It covers half of the observations such that one fourth of the remaining data lies beyond either end of the box. The bold bar inside the box represents the median, i.e. the second quartile. Dashed lines reach to the most extreme data point which is no more than 1.5 times the interquartile range away from the upper or lower box edge.

3 Results and Discussion

3.1 Secondary structure sustains random evolution

We monitored the behavior of predicted secondary structure under in-silico mutation (s. section 2.1). Our analysis revealed an unexpected high robustness of secondary structure content and segment lengths towards our mutation protocol. More specifically, with ongoing sequence divergence, the averages in content and length of predicted helices and strands stayed at an overall constant level. Even at low levels of sequence identity, helix content remained at ~30% and helix length at ~10 residues (corresponding to 2-3 helix turns) (Fig. 3.1A,B: green box plots).

For predicted strands, we observed the same trend. In sequences similar to random, average strand content was at ~20% while strand length remained at a level of around five residues (Fig. 3.1C,D: green box plots). These values were nearly identical to those found in native sequences.

The stability in average content and segment length was independent of the chosen mutation scheme (background versus PAM120), although we observed an insignificant decrease in helix content and length during mutation according to PAM (Fig. 3.1A,B: yellow box plots). Identical levels of distributions between very low sequence identity and random sequences (Fig. 3.1A-D: two rightmost green box plots) suggested that we mutated long enough to loose any 'memory' from native sequences.

We also addressed two potential deficits in our experimental setup. First, since no structural information of random amino acid sequences exists in large scale, we were forced to base our investigations on predictions instead of observations. However, our predictions may have been prone to mistakes, especially due to the circumstance that we explicitly had to use PROFsec in sequence-mode. The only place where we were able to shed light on whether our findings were in fact prediction artifacts was before we started the in-silico mutation. In native PDB sequences we compared secondary structure predictions with observations (derived from DSSP, s. section 2.1). For both, our content and length distributions were indeed the same, except for an insignificant over-prediction of strand length (Fig. 3.1A-D: light and dark gray bars). This result suggested that errors in predictions did not matter for our coarse-grained measures of change.

Second, our findings may have been biased by either under-representation or

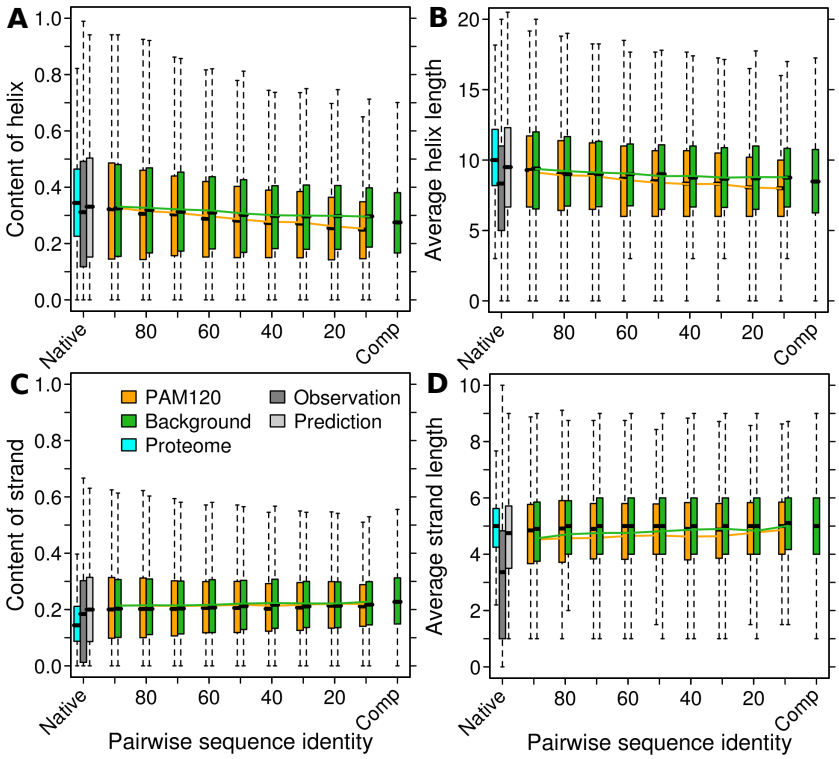

Figure 3.1: Secondary structure content and length stable. We monitored change in content and length distributions of predicted helix (**A,B**) and strand (**C,D**) during mutation. Box plots denote spread in distributions (section 2.5). Green and yellow bars depict mutation according to background amino acid composition and PAM120, respectively. Dark and light gray box plots denote observations and predictions in native sequences, blue bars predictions in human proteome. Right-most green bars (labeled 'Comp') represent predictions in random sequences. Overall, neither the length nor the content of regular secondary structure appears to differ between native and random. (Adapted from Schaefer *et al.* (2010))

complete absence of certain protein families in the underlying sequence database (PDB). To resolve this potential shortcoming, we furthermore compared predictions between wild-type sequences and proteins representing an entire proteome. Our results suggested, that neither content nor length of both helices and strands differed between two sequence sets, in fact distributions were virtually identical (Fig. 3.1A-D: blue vs. light gray box plots). Based on this, we concluded that potential bias did not affect our findings.

Overall, we observed a surprisingly high robustness of secondary structure against mutations by our coarse-grained measures, that is, the constant upkeep of its con-

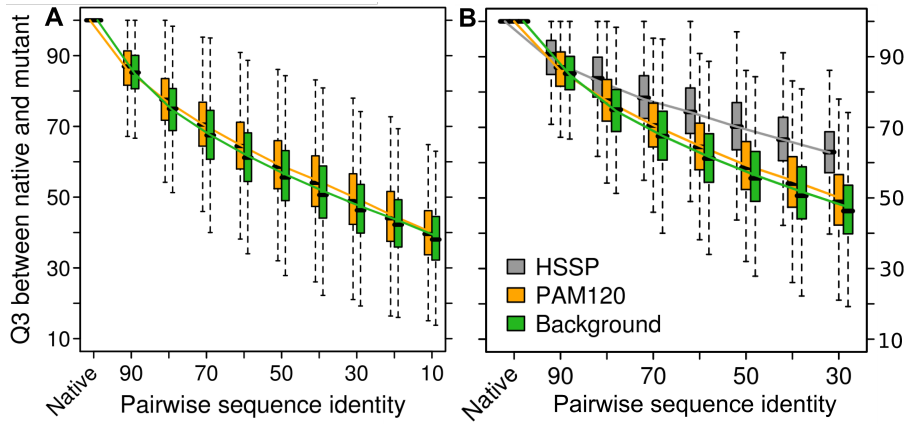

Figure 3.2: Secondary structure diverges linearly to sequence changes. We monitored change in secondary structure (Q3) dependent on sequential change (pairwise sequence identity). During our in-silico mutation (**A**, yellow and green box plots, s. section 2.5), secondary structure diverged linearly to an random level of ~33% Q3, while pairs of naturally evolved homologues (**B**, gray box plots) showed less divergence in secondary structure: At 30% sequence identity, natural homologues still showed a Q3 ~63% while random mutants reached ~45%. This difference can be attributed to the enrichment of structural neutral mutations under evolutionary constraints. (Adapted from Schaefer *et al.* (2010))

tent and length. Earlier studies showed that with increasing divergence of two naturally evolved protein sequences their 3D structures (Abagyan and Batalov, 1997; Chung and Subbiah, 1996; Sander and Schneider, 1991) and secondary structure in particular (Rost *et al.*, 1994, 1997) also become increasingly different.

Our analysis revealed a similar result for the artificially created mutant sequences. During the course of the in-silico mutation, we monitored the difference of predicted secondary structure between native and diverging sequences in terms of Q3 measure (fraction of residues identical in either one of three states helix, strand, other between wild type and mutant sequence). The rate at which secondary structure changed was nearly linear to changes in sequence (Fig. 3.2A). In random-like sequences, we observed an average Q3 of ~33% which resembles the probability of picking a particular state (of either helix, strand or other) at random. Put differently, while sequences diverged to random levels so did their secondary structure. This behavior was independent of the mutation scheme (Fig. 3.2A, PAM120 vs. background).

To relate this finding to naturally evolved homolgues, we conducted this analysis also on protein pairs at different levels of sequence identity, taken from the HSSP database (Sander and Schneider, 1991). The rate at which secondary structure

diverged was less pronounced as compared to our in-silico mutation (Fig. 3.2B, gray vs. yellow/green box plots), which confirms the expected: The exposure of naturally evolved proteins to selective pressure - which was not built in our mutational model - led to an enrichment of neutral mutations with respect to structural change.

Overall we found that neither the content nor the length of predicted secondary structure differed between native proteins and random amino acid sequences. This suggested that the formation of secondary structure is a property that is inherent to amino acid sequences and that its upkeep during evolution might not be too challenging. This finding may contribute to the perception that protein structure is quite robust against sequence changes and that a variety of sequences fold into the same structure (section 1.2).

Nonetheless, despite this high robustness of content and length, the specific secondary structure states in highly diverged sequences did not resemble those in native proteins. Both circumstances, the constance in content and length plus a Q3 of 33%, suggested a rather random-like concatenation of secondary structure segments in random amino acid sequences.

3.2 Structural effects are in the details

Helices become strands and vice versa

Our analysis showed that the formation of secondary structure is not an exclusive property of naturally evolved proteins. It rather appears to be *intrinsic* to any (random) amino acid sequence. Nonetheless, as sequences diverged from their native state during our mutation protocol so did their secondary structure.

To investigate change in detail, we analyzed structural effects in mutation trajectories of four proteins representing the four main SCOP (*Structural Classification of Proteins*, Murzin *et al.*, 1995) classes (Fig. 3.3). Two major observations stood out.

First, regions containing regular secondary structure elements constantly transited from one state to the other, that is, helices suddenly became strands and vice versa. This behavior occurred more often than transitions from helix to coil or strand to coil. Second, while the ends of these regions continuously shortened and extended, their core regions remained overall robust. This led to the upkeep of stable blocks consisting of interchanging regular secondary structure throughout the 69 mutation steps. Nonetheless, in the end almost no native helix or strand withstood our mutation protocol.

Our observations of the dynamics during in-silico mutation substantiated the previous findings in that secondary structure changes with ongoing sequence di-

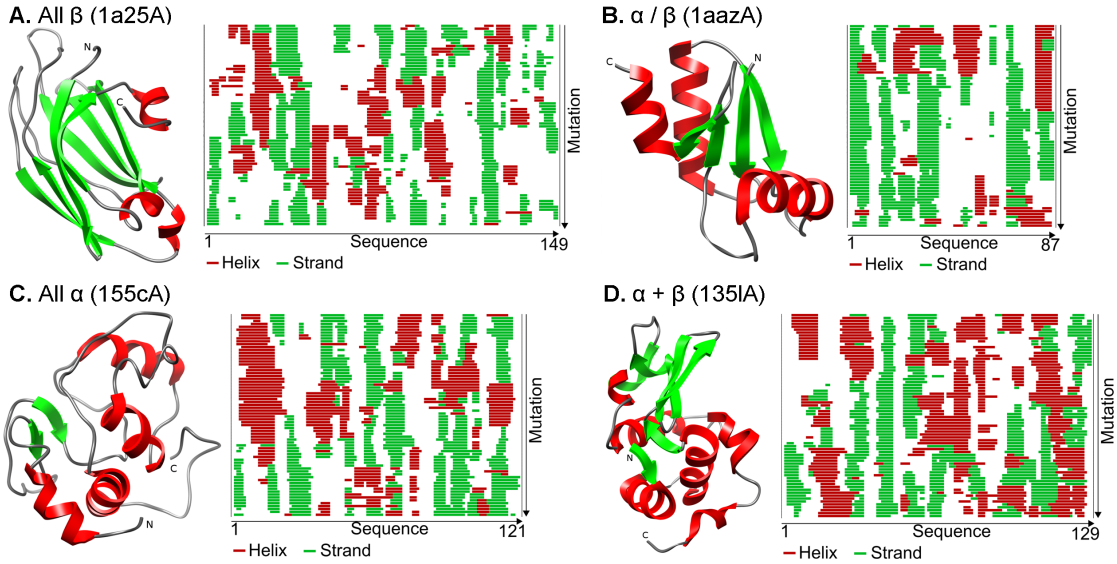

Figure 3.3: Secondary structure states switch back and forth. We picked one representative for each of the major four SCOP classes (Murzin *et al.*, 1995). (**A-D**) We depicted each protein through its ribbon plot and mutation trajectory, consisting of the native sequence on top and 69 mutated sequences with ongoing levels of divergence below. Sequences were mutated according to PAM120. Green and red regions in ribbon plots and trajectories denote strand and helix regions, respectively. Two observations stand out. Secondary structural elements flip back and forth during our mutation protocol, while blocks of regular structure remain quite robust. Structures were rendered with Chimera (Pettersen *et al.*, 2004). (Adapted from Schaefer *et al.* (2010))

vergence. Small sequential changes could lead to local conformational impacts induced by interchanging secondary structure elements. Recent experimental work revealed a much more dramatic change upon a single point mutation between two structural distinct classes (mainly alpha to mainly beta) (Alexander *et al.*, 2009).

Short disorder comes and goes, long disorder goes

We subjected predicted disorder to a similar analysis to investigate its resilience against our mutation protocol. Protein disorder displays two different occurrences, that is, very short and very long regions (Dosztanyi *et al.*, 2005b; Liu *et al.*, 2002; Obradovic *et al.*, 2005; Schlessinger *et al.*, 2009). A common practice in the field is to apply strict length thresholds to distinguish between both, although these are not biophysically substantiated. They rather strive to exclude the ambiguous region in between that resembles characteristics of both regimes. We adhered

to these conventions and considered regions below eight consecutive disordered residues as short disorder and regions above 30 consecutive disordered residues as long disorder.

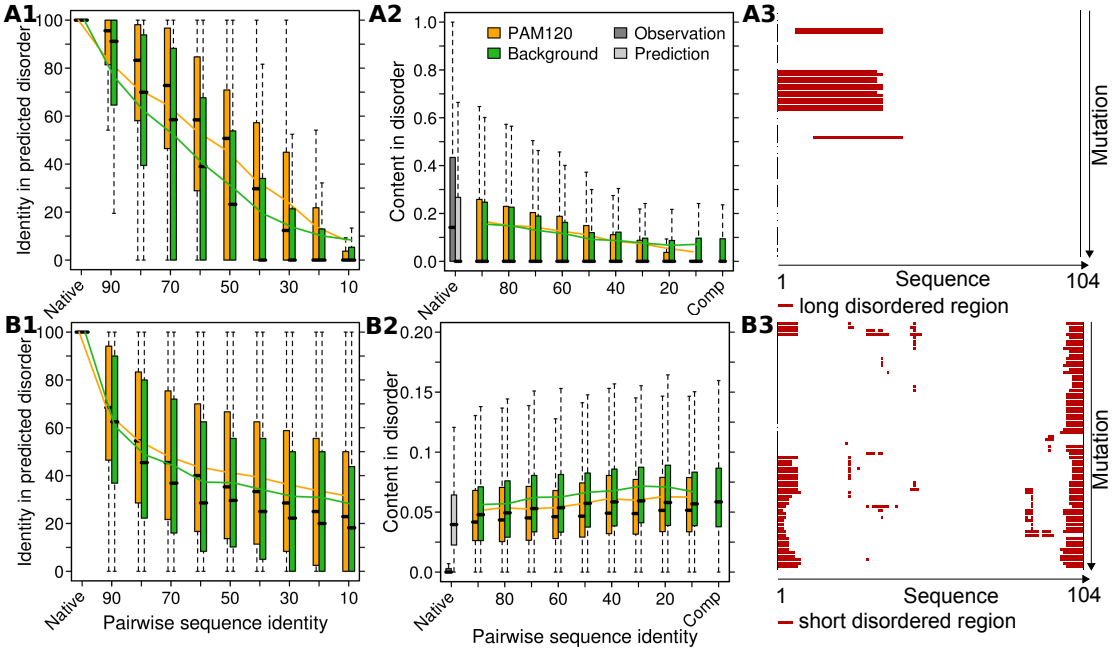

Figure 3.4: Long disorder disappears, short disorder fluctuates. The top row shows the behavior of long disorder under random mutation, the bottom row that of short disorder. Predictions for long disorder (**A1**) drastically diverge (measured by identity in predicted long disorder between native and mutated sequence, y-axis) from native states with ongoing mutation (x-axis); predicted short disorder (**B1**) diverges more slowly. (**A2, B2**) Dark and light gray box plots compare disorder content in observations with that in predictions; rightmost green box plots denote distributions in random sequences (labeled 'Comp') and ensure convergence during mutation. (**A2**) While content in predicted long disorder tends to disappear upon mutation, (**B2**) that of short disorder remains constant. (**A3, B3**) Mutation trajectories in a representative example (DisProt identifier DP 00006) with predicted long (**A3**) and short (**B3**) disorder show the wild type sequence on top followed by 69 mutants according to PAM120. Apparently, long disordered regions disappear while short disorder re-/disappears. (Adapted from Schaefer *et al.* (2010))

For short disorder, we observed a similar picture as for secondary structure, that is, both its average content and length remained stable during the mutation protocol (Fig. 3.4B2,B3 for content; length distributions not shown here, s. Schaefer *et al.* (2010)) while it gradually diverged from the native states, as measured by

Q2 (Fig. 3.4B1). More specifically, we observed ~5% of predicted short disorder in native sequences which slightly increased to ~6% in random sequences.

Long disorder exhibited a different behavior: it tended to disappear upon our in-silico mutation protocol (Fig. 3.4A1-A3). Its content decreased from ~18% down to 2-9%, depending on the mutation scheme (Fig. 3.4A2, yellow vs. green box plots). The disappearance of long disorder was more pronounced in mutation according to PAM120 than to background. Since PAM substitutions were expected to 'push' disordered proteins to an amino acid composition resembling that in ordered proteins, this was not unexpected. The analysis of the mutation trajectories of two representatives confirmed our findings. Long disordered regions vanished after half of the mutation procedure (Fig. 3.4A3), while for short disorder we observed a constant dis- and reappearance (Fig. 3.4B3), which was specifically apparent at both protein termini.

The loss of disorder induced by little sequence variations may have further phenotypic consequences. Disordered regions play an essential role in protein function (Dunker *et al.*, 2002; Dyson and Wright, 2005; Vucetic *et al.*, 2007). Recent work suggested that point mutations in disordered regions are often linked with disease (Hu *et al.*, 2011; Ye *et al.*, 2007) and that the transition to well-ordered structure may be reason for this effect (Vacic and Iakoucheva, 2012).

Our analyses showed that the content of short disordered regions in disordered proteins remained stable. In that respect, it acted similarly to secondary structure (Fig. 3.3). Another aspect of short disorder became apparent: Short disorder often was predicted at both protein termini (Fig. 3.4B3), a feature especially observed in loopy ends of globular proteins due to their lack regular secondary structure in these regions (Liu *et al.*, 2002). On the other hand, the volatility of long disorder to random mutation makes this feature in disordered proteins a feature *not intrinsic* to amino acid sequences. It is therefore likely that long disorder needs to be actively maintained during evolution by selection against mutations disrupting it.

3.3 Structural change predictable from sequence

Our coarse-grained analyses of structural change under in-silico mutation showed: Changing the amino acids in 10% of a sequence could lead to local effects pertaining to switches between different secondary structure states and the formation or disappearance of disordered regions. These measures of structural impact upon sequence change however were rather coarse grained. Furthermore, our analysis did not link one particular amino acid change with its individual effect it has on structure.

The implications of single amino acid exchanges on protein structure have been studied before. Early investigations on 83 X-ray mutant structures in the PDB

(Berman *et al.*, 2000) led to a set of predictive rules based on position-dependent rotamers (De Filippis *et al.*, 1994). The lack of structural variety contained in that small set of proteins however makes it unclear how well such a method would perform in the diversity of structures in the contemporary PDB.

Therefore, we approached the objective of predicting structural change upon a point mutation differently. We compiled a set of pairs consisting of two structurally superimposed pentapeptides. Each pair had two different amino acids in its center while the flanking regions were identical in sequence. In addition, each such pair was labeled either *structural neutral* or *non-neutral* and had an associated ground set of sequence derived features. To ensure a realistic performance assessment, we conducted a feature selection procedure and the subsequent performance estimation on two separate subsets of pentamer pairs (s. section 2.2).

Three features most predictive

During our forward selection procedure (s. section 2.2.4), we found the following three most informative features: difference in PSIC values between native and mutant residue, predicted three-states secondary structure (raw values from PROFsec, window around mutant w=17) and BLAST information per position (w=21). These properties already raised the predictive performance of our model on the holdout set to an AUC of ~0.82.

Six further features were added to the list until no performance gain of ΔAUC > 0.001 was observed. The following properties raised the overall performance only marginally by 0.02: predicted residue flexibility (raw output from PFOFbval, w=21), differences in PSSM and predicted secondary structure between native and mutant residue, HMMER scores for fitting amino acids into a PFam domain alignment (w=13), predicted protein-protein interaction sites (raw values from ISIS, w=13) and the amino acid volume (w=5).

All features led to an average AUC of ~0.84 during 10-fold cross validation on the holdout set. Due to the specific encoding of these properties, the overall feature space contained 147 numerical feature values.

Model predicts structural change accurately

We assessed the predictive performance of our logistic regression model during a 10-fold cross validation, conducted on pentamer pairs not used during feature selection (s. section 2.2.4). The model returned a probability estimate p for structural change. By applying the default cutoff of 0.5, an amino acid change was assigned a binary prediction of either *structural change* ($p > 0.5$) or *neutral* ($p \leq 0.5$). By iterating over all such probability thresholds, we established a ROC curve (Fig. 3.5A) and accuracy-vs-coverage plots (Fig. 3.5B) (s. section 2.2.5).

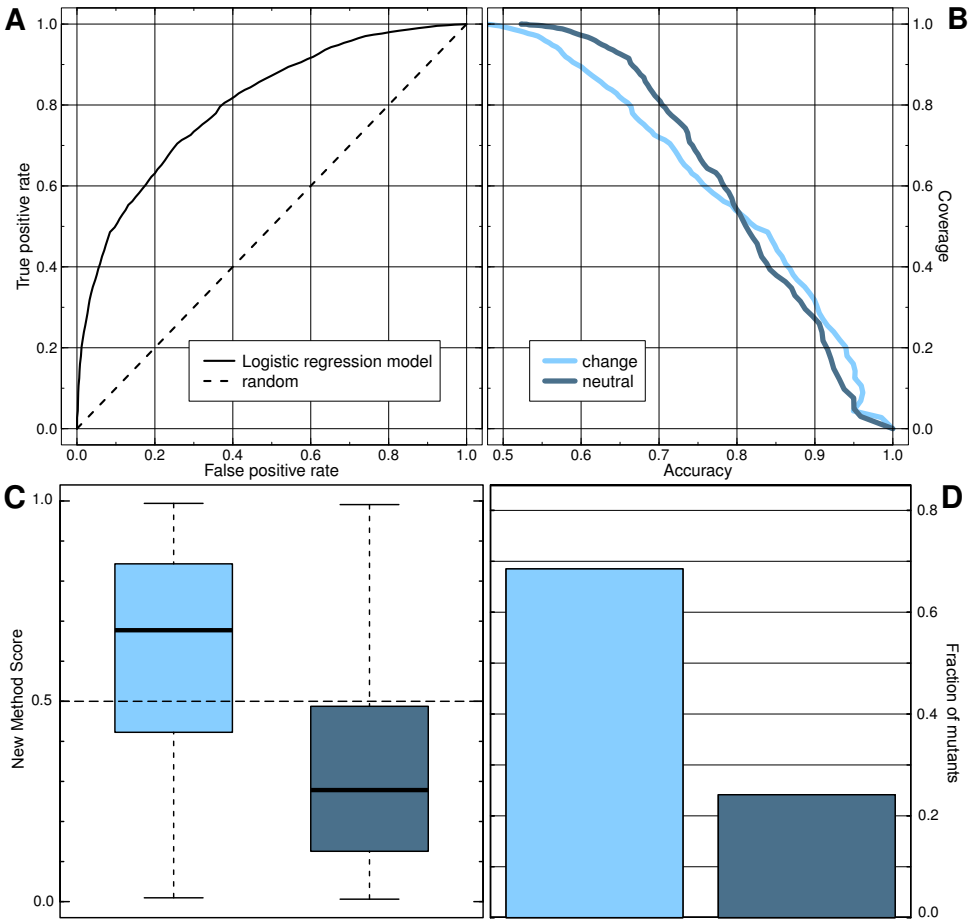

Figure 3.5: Structural effect predictable from sequence. Different performance statistics conducted during a 10-fold cross validation on a separate set not used during feature selection. (**A**) The ROC-curve of our method (solid line) suggested an overall AUC ~0.8 compared to an AUC of 0.5 for random predictions (dashed line). (**B-D**) Good discrimination between the two classes structural change and neutral; separate accuracy-vs-coverage plots (**B**) revealed similar performances for change (light blue) and neutral (dark blue) predictions. (**C**) Box plots (section 2.5) of method scores demonstrated high separation between two classes, i.e. higher scores (>0.5, dashed horizontal line) were more prevalent in effect predictions (light blue), lower scores (<0.5) more prevalent in neutral (dark blue); (**D**) at default threshold of 0.5 the method predicted ~69% of structural change predictions were correct (true positives), while only ~23% of structural neutral were predicted falsely (false positives). (Adapted from Schaefer and Rost (2012))

The final model reached an average AUC of ~0.8 and an overall two-state accuracy Q_2 of ~72% after applying the default threshold of 0.5. These measures reflected the overall performance without revealing separate class behaviors. The accuracy-vs-coverage plots (Fig. 3.5B) revealed that ~52% of neutral and effect predictions reached an accuracy of ~80%. For higher accuracy, the correct predictions were dominated by predictions of change. This high performance was also evident in a high separation of prediction scores p between both classes: Larger scores ($p > 0.5$) were much more abundant in the class *change* while small scores ($p \leq 0.5$) were dominant in the *neutral* class (Fig. 3.5C). At the default threshold of 0.5, ~69% of structural change instances were predicted as such, while only ~23% of structural neutral were predicted as change (Fig. 3.5D).

These results suggested that sequence-derived information sufficed to predict structural change upon single amino exchange in pentamers. This was especially remarkable due to the circumstance that structural conformations of pentapeptides crucially depend on their specific structural neighborhood (Cerpa *et al.*, 1996; Fliess *et al.*, 2002; Kabsch and Sander, 1984). Explicit knowledge about that, however, was not included in our input features.

3.4 Observed effects enriched in predicted structural effect

Strictly speaking, our method learned how to separate to different populations of peptide pairs that consisted either of two structurally similar pentamers or of two structurally dissimilar ones. We expressed structural dissimilarity in terms of a conformational shift between both peptide backbones, measured by RMSD over C_α atoms (s. section 2.2.1). The underlying motivation however was to predict structural effects induced by single amino acid exchanges in proteins and to use this knowledge to gain deeper insights into other biologically relevant effects.

The specific molecular details inherent in a structure determine other aspects such as the stability or the function of a protein. Changes in structure, e.g. by exchanging an amino acid for another, could lead to further consequences pertaining to stability and function with possible consequences on even phenotypic level such as a raised susceptibility to disease (Wang and Moult, 2003, 2001; Gong and Blundell, 2010; Stitziel *et al.*, 2003; Sunyaev *et al.*, 2000; Talavera *et al.*, 2010).

However, these implications are not strict or do not apply in both directions. On the one hand, it is evident that mutations leading to effects on stability or function are expected to be enriched in those that severely alter structure. On the other hand, the absence of a structural effect measured in terms of a backbone shift does not necessarily imply an unaltered function. For example, changing the

intricate details of hydrogen bonds donors and acceptors in active sites through different side chain conformations may inactivate enzymatic function but leave the backbone rather unaltered. Nonetheless, strong structural effects are expected to raise the likelihood for other effects.

If the method had learned important aspects about structural change beyond the originally posed task, it should contain an intrinsic ability to reveal an enrichment of experimentally observed effects in strong predicted structural effect. We tested this hypothesis on three different sets of annotated single amino acid exchanges. The first set contained mutations that do or do not alter protein stability, the second set comprised functionally neutral and non-neutral polymorphisms and the third set consisted of nsSNPs that had either a disease annotation or were non-deleterious (s. section 2.3.1).

We monitored the fractions of observed effect mutants in those that had a predicted structural effect and compared these to the background distributions found in the datasets. We defined predicted structural effect at two different probability thresholds as returned by our method. First, we considered each prediction at the default cutoff of $p > 0.5$ as structural effect and second, we used a higher threshold of $p > 0.9$ reflecting strong structural effect.

One major result stood out. Compared to the background, mutations with an observed effect on any of the three were enriched in those that were predicted to have an effect on structure (Fig. 3.6). Furthermore, the observed accumulation increased with the severity of the predicted effect. We observed the most pronounced signal in disease-annotated mutants: their fraction increased from 39% in background over 48% in structural effect to 68% in strong structural effect which translated into an overall enrichment of 29% (Fig. 3.6: three rightmost bars). Functional effect exhibited a similar strong increase in enrichment of overall 26% from background (49%) to strong structural effect (75%) (Fig. 3.6: middle bars). The accumulation of stability-changing mutants was least pronounced with 14% (Fig. 3.6: leftmost bars) and little significant due small sample sizes.

These findings strongly suggested that our method not only succeeded in separating structural neutral from non-neutral pentamer pairs. More importantly, they also showed that our rather artificial definition of effect captured indeed important aspects of structural change upon point mutation. Nonetheless, a significant fraction in strongly predicted structural effect mutations *did not* exhibit a disease consequence (i.e. 32%) or a functional effect (25%). Since it is not clear whether those are truly neutral or not yet experimentally annotated as non-neutral, we had no means to attribute this discrepancy to either a true signal or to deficiencies in our methodology.

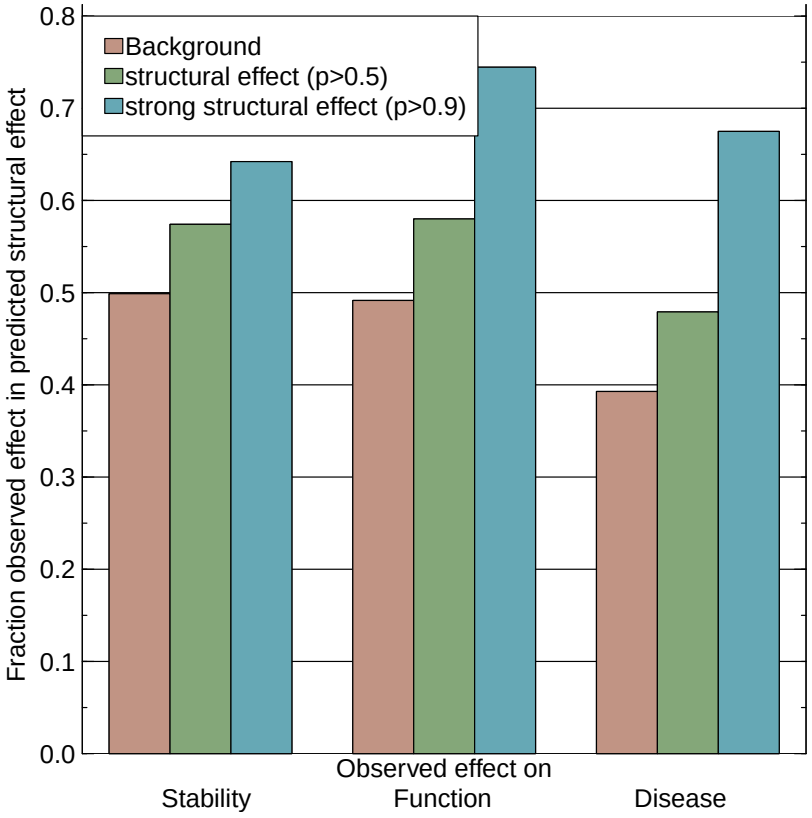

Figure 3.6: Mutations predicted to affect structure are often linked with disease and change in function. We considered mutations predicted as changing structure moderately ($p > 0.5$) and severely ($p > 0.9$). Mutations having an observed effect on stability (left), function (middle) and disease (right) occurred more often than expected from background (brown bars). Their enrichment increased with increasing severity in predicted structural effect mutants (green and blue bars). This suggests that strong structural impact upon single amino acid exchange increases the likelihood of other effects such as disease.

3.5 Disease strongly correlated with predicted functional effect

We established that disease-related mutations occur more often in mutations that are predicted to severely alter structure. We tested a similar hypothesis for function-changing mutations and asked whether a strong functional effect is correlated with disease. We predicted the functional severity through SNAP (Bromberg and Rost, 2007) of 26,404 mutations with disease relation but no observed effect and compared them with 251,414 variants that had no disease annotation (s. sec-

tion 2.3.2).

The predictions differed greatly between both sets in two respects. First, SNAP predicted much more disease-related mutations to change function than mutations with unknown disease relation. More specifically, at the default threshold of 0, SNAP revealed ~86% of disease-related variants as functionally non-neutral while only ~51% in the set of unknown disease relation were predicted to change function (Fig. 3.7A,B: black dashed vs. green curve). Second, functional change was predicted to be stronger in disease-related mutants than in unknown disease mutations. This finding was reflected by a more pronounced shift of SNAP scores towards larger values in disease related mutations. About 47% of these variants exhibited scores of more than 40 compared to only 12% of mutations with unknown disease relation at that score (Fig. 3.7A).

The magnitude of predicted effects in disease-related mutants became even more pronounced through a comparison with functional predictions in variants with an observed effect but without known disease association (s. section 2.3.2). These mutations constituted the functional non-neutral part of the data used to train SNAP. As any other machine learning based method, SNAP performs significantly better on its own training set than on data that did not participate during its optimization. As a result, scores in this set are expected to be biased towards 'more effect' compared to predictions in any other set of effect mutations not used for training.

We observed a quite different outcome. Only 40% of the training set was predicted at scores >40 which was 7% less compared to disease related mutants (Fig. 3.7B: dashed black vs. red curve). For variants annotated to be disease-related *and* having an observed effect, SNAP revealed the highest effect on function: At the default threshold, about 90% were predicted to have a functional impact (more than 4% than in disease related) and ~53% had scores higher than 40, i.e. 6% more than in disease related mutations (Fig. 3.7B: solid black curve). This suggested that disease-related mutations alter protein function more severely than any other mutation with an observed effect.

In this context, we also addressed the question of how many potential disease associations are yet undiscovered. A considerable amount of variants with unknown disease relation was predicted to have a slight (51%) or severe effect (12%, SNAP score >40). We considered the amount of effect predictions in random mutations that occurred maximally distant to any observed mutation (s. section 2.3.2) as background. This enabled us to estimate an upper bound of undiscovered associations in this set. SNAP predicted 7% of random mutations as severely changing function (Fig. 3.7B: blue curve). Comparing this fraction of high impact mutants to the one predicted in variants without disease annotations suggests that 7% - 14% still remain to be experimentally annotated as severely altering function and

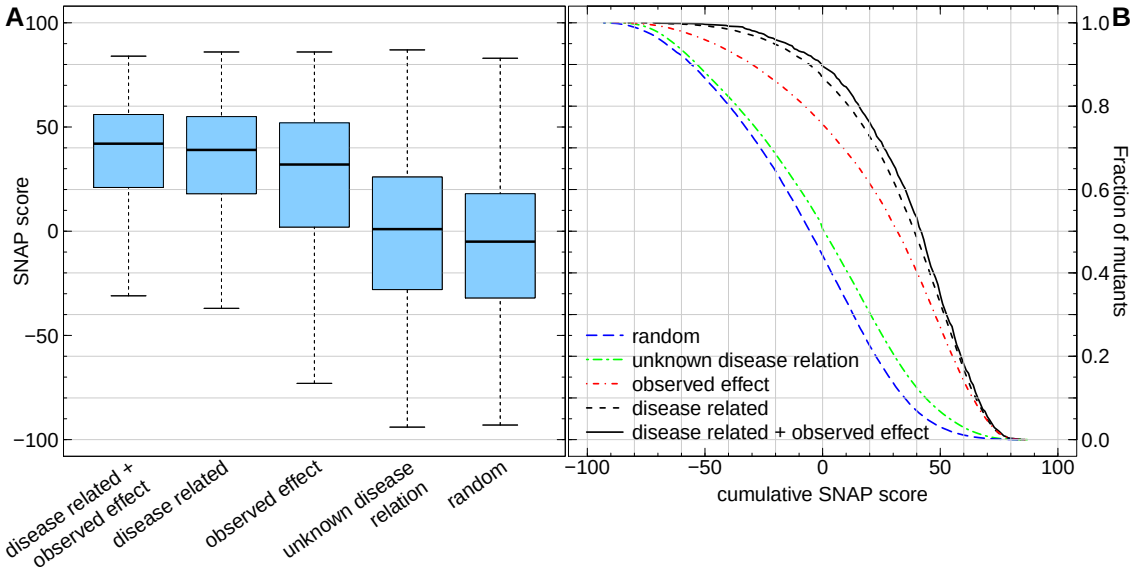

Figure 3.7: Disease-related mutations predicted to severely alter protein function. We predicted the severity of functional impact in five different sets of point mutations using SNAP: *disease related + observed effect*, *disease related* (without observed effect), *observed effect* (without known disease relation), *unknown disease relation* and *random* mutations. (**A**) Box plots (s. section 2.5) depict the distributions of SNAP scores in all five sets, the distance from 0 denotes the severity of effect; the fraction of mutants above the default threshold of 0 are predicted as non-neutral, above 40 as high impact mutations. Disease causing mutations contained the highest fractions of functional non-neutral variants (90% and 86%, two left box plots), while impact predictions dominated *observed effect* mutants even less (76%, middle box plot). Predicted effect in random mutations (44%, rightmost box plot) provided an upper bound for effect variants in those with *unknown disease relation* (51%). (**B**) Cumulative predicted severity; points on a curve correspond to fractions (y-axis) of mutations with severity (x-axis) $\geq$ that score. Disease causing variants (black solid and dashed curve above all others) were predicted to have the most severe impact on protein function. (Adapted from Schaefer *et al.* (2012a))

thus as candidates for disease causing variants.

Through predictions of functional effects in mutations with and without disease relation, we correlated severe impact on protein function with disease. Put differently, if a mutation leads to a disease then a strong functional change may play a major role in explaining its reason. Despite the abundance of high impact mutations in disease causing variants, we observed nonetheless a pronounced overlap of score distributions between disease related, unknown disease relation and random

mutations (Fig. 3.7A). Thus our investigation did not shed light on the reverse, i.e. whether a strong functional change implies disease.

3.6 Functional change predicts disease accurately

Our findings so far led to the following key indication: Disease mutations are enriched in predicted effects on either structure or function (sections 3.4 and 3.5). These findings suggested that a strong molecular impact on protein level appears to raise the likelihood for disease, as also reported by others (Gong and Blundell, 2010; Stitziel *et al.*, 2003; Sunyaev *et al.*, 2000; Talavera *et al.*, 2010; Wang and Moult, 2003, 2001). Ultimately this raised the following question: Could the knowledge contained in methods that predict different effects on molecular level be combined to predict disease at higher confidence than each method could do on its own?

To investigate this question, we predicted structural and functional effects in disease related and non-deleterious mutations (s. section 2.3.1) through the new method developed here and SNAP, respectively.

At the default threshold of $p > 0.5$, the new method predicted ~81% of disease-related mutations as affecting structure while ~57% of non-deleterious mutations were predicted as such (Fig. 3.8A: fractions above 0.5 of two leftmost box plots). Strong structural effect ($p > 0.9$) was predicted for ~22% of disease-related and ~7% for non-deleterious variants. These results confirmed our previous findings in that predicted structural effect is enriched in disease associated mutations (s. section 3.4). However, our method predicted a significant amount of strong structural effect in non-deleterious mutations and, vice versa, a significant amount of disease-related mutations were predicted as structurally neutral (Fig. 3.8A: strong overlap of two leftmost box plots).

Furthermore, we assessed the predictive performance in detail and separately for disease-related and neutral mutations through ROC-like curves. This approach allowed us to monitor accuracy-coverage pairs independent from a default decision threshold by iterating over the whole range of prediction scores (s. section 2.2.5). First, we observed that at the default cutoff of $p > 0.5$, the accuracy for neutral predictions was as high as ~78% but only 42% of non-deleterious mutants achieved that level of accuracy (Fig. 3.8B: solid and dashed green curves). The observed accuracy for disease predictions at that cutoff was only ~47% and more than 80% reached that performance level (Fig. 3.8B: solid and dashed brown curves). For very strong predictions, i.e. at a threshold of $p > 0.88$, the method achieved a balanced accuracy of 66% for both classes (Fig. 3.8B: arrow). However, only 24% of disease-related but >90% of neutral mutations were predicted with that level of accuracy. Our analysis showed that no threshold existed to achieve a

high performance for predicting a large amount of disease and neutral mutations correctly. Structural effect as predicted by our method did not suffice to clearly separate disease from neutral.

We observed a different behavior in predicted functional effect. At the default threshold of 0, SNAP predicted 85% of disease-related mutations as functional non-neutral but only 20% in non-deleterious variants (Fig. 3.8A: fractions above 0 of two rightmost box plots). At a cutoff of 40, 46% of disease-related mutations were predicted as strongly altering function while only 4% of non-deleterious were predicted as such. The overlap between both distributions of predicted functional severities was by far less significant compared to the predictions of structural change (Fig. 3.8A).

We conducted the same detailed performance analysis detached from a specific SNAP score threshold as in predictions of structural change. At the default threshold of 0, SNAP predicted 80% of non-deleterious with an accuracy of ~90% and ~85% of disease-related mutations with an accuracy of 73% (Fig. 3.8C: dashed and solid green curves). We observed a balanced accuracy of 81% at a cutoff of 22. At this threshold, 90% of non-deleterious and 70% of disease-related mutations were predicted having that performance (Fig. 3.8C: arrow).

Apparently, functional change as predicted by SNAP led to far better criterion to separate disease-associated from non-deleterious variants than predicted structural effect. This finding however has to be taken with a certain grain of salt. Part of the disease-related mutations with an additional observed effect as well as the non-deleterious variants took part in SNAP's training (section 2.3.1). The accuracies reported here should therefore be considered as biased towards better performance. However, this positive effect was marginal as SNAP performed only slightly worse during its cross validation (~80% accuracy/coverage for both neutral/non-neutral at SNAP score 0 (Bromberg and Rost, 2007)).

Could the combined prediction of both effects increase the performance of disease prediction? This should have turned out to be the case, if our new method contained knowledge about protein structure relevant for predicting disease which, in addition, was orthogonal to the information about function intrinsic in SNAP. To test this hypothesis, we considered a prediction as deleterious if SNAP returned a score of 22 (i.e. the threshold SNAP performed best at distinguishing disease from neutral, s. above) and the new method predicted a structural effect. We again iterated over the whole range of prediction scores for structural change and monitored pairs of accuracy and coverage for disease and non-deleterious predictions.

The accuracy of disease predictions did hardly increase over the entire spectrum of structural change cutoffs (Fig. 3.8D: brown solid curve). Only at thresholds $p > 0.9$ the accuracy started to raise above 80%, however only for <20% of disease-related mutations (Fig. 3.8D: arrow). We observed a similar performance in SNAP-

only predictions at scores >22 but with a much higher coverage of ~60% (Fig. 3.8C: solid/dashed brown curves at regions right of arrow). For non-deleterious variants, the observation was similar: At a cutoff > 0.9, >90% of neutral mutations reached an accuracy of 65%. However, SNAP alone topped this performance again: At scores >22, it predicted >90% non-deleterious variants with an accuracy of 80% (Fig. 3.8C: solid/dashed green curves at regions right of arrow).

Our analysis showed that the combination of predicted effects in both structure and function did not raise the performance to predict disease. Despite our finding that strong structural change is enriched in disease variants, our method predicted a significant amount of structural change in non-deleterious mutations and a lot of disease mutations as structurally neutral.

Our investigations did not address the question whether anything is special about these outliers in a biological sense. It also could very well be the case that our definition of structural change did not capture essential knowledge about the intricate details in protein structure that, when being changed upon mutation, lead to disease. Thus, further investigations will have assess the capabilities of other potential measures for structural impact. Possible candidates could be changes in H-bond donors/acceptors (s. section 1.3 and Gong and Blundell (2010); Wang and Moult (2001)) or side chain torsion angles.

The information alone that a mutation triggers functional change apparently sufficed to predict disease at high accuracy. In the chain of causal relationships that connects basic molecular effects on its one end and phenotypic changes on its other, protein function appears to play a role as key link.

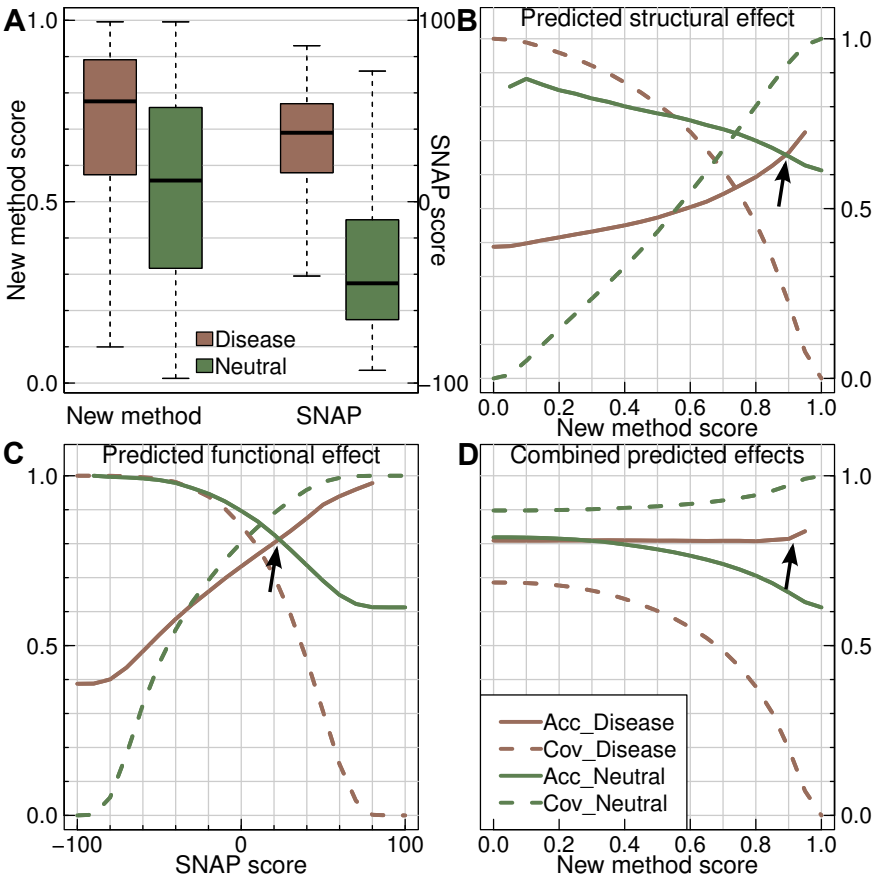

Figure 3.8: Functional change alone predicts disease accurately. We predicted structural and functional change in disease-related and non-deleterious mutations through the method developed in this thesis and SNAP, respectively. (**A**) Box plots (section 2.5) show distributions of method scores for predicted structural (new method, left pane) and functional effect (SNAP, right pane) separately for disease (brown) and neutral (green) mutations. Strong structural and functional effect is predicted more often in disease-related than in neutral mutations. Predicted functional effect separates disease-related better from neutral mutations than predicted structural effect (little vs. much overlap between brown and green box plots). (**B-D**) ROC-like curves depict accuracy-coverage pairs (section 2.2.5) sampled at different prediction cutoffs of SNAP and the new method. Our new method achieved the best performance at distinguishing disease-related from neutral mutations for p>0.88 (**B**, arrow), but only few disease-related mutations were predicted at that cutoff. SNAP performed better, i.e. at a cutoff of 22 the majority of disease-related and neutral mutations were predicted at high accuracy (**C**, arrow). We combined both methods by considering a deleterious prediction as both functional (SNAP score >22) and structural non-neutral (sampled over entire cutoff range); the overall performance did not profit, i.e. for very few disease-related mutations, we observed a slight increase in accuracy for deleterious predictions at very strongly predicted structural change (**D**, arrow).

4 Conclusion

The scope of this thesis was the prediction of structural effects that occur upon a residue exchange in proteins. Of particular interest was the question whether a mutation predicted to impact protein structure was likely to be involved in disease development. In other words, could the knowledge about a structural change aid in predicting disease?

In a first assessment, we perceived structural change in a rather coarse-grained way. Specifically, we tackled the question as to how predicted secondary structure and protein disorder changed under random mutations. Ultimately, this also shed light on how easy it is for evolution to maintain these structural features. Our findings clearly suggested two different implications. First, neither the content nor the length of helices and strands changed significantly with ongoing sequence divergence. In stark contrast, long disorder was disrupted during random mutation and disappeared in random sequences. Hence, it appeared very likely that the formation of well-ordered secondary structure is intrinsic to any amino acid sequence. The upkeep of long disordered regions over the course of evolution appears to be more challenging. Second, despite the high robustness in content and length, we observed ongoing transitions between helices and strands as well as a continual coming and going of short disordered regions. We found that small sequence changes affected local structure significantly.

In a next step, we introduced a finer-grained definition of structural effect induced by a single residue exchange. We perceived structural change as strongly displaced backbones of two protein fragments which differed only in their central amino acid. We attributed structural impact to the residue exchange. This approach allowed for the compilation of a large dataset and the successful training of a machine-learning method. Its objective was to separate structural neutral from non-neutral fragments based on sequence-derived features.

Through predicting structural effects of mutations that exhibited observed effects, we were able to show that the new method captured important biological aspects beyond our rather artificial definition of change: We established that effects on protein stability and function were enriched in mutations predicted to have a strong structural effect. Even more importantly, we related a strong structural effect to an increased likelihood of a mutation to be disease-related. Similarly, we found that a method trained to predict functional effect clearly identified disease-related mutations as severely altering protein function.

Our findings indicated that methods trained to predict molecular effects could provide a valuable step towards predicting the deleteriousness of point mutations. We tested this hypothesis and used predictions of structural and functional effects on their own and in combination to distinguish disease-related from non-deleterious variants. The outcome was clear in that predicted functional method alone sufficed to accurately predict disease.

Future investigations will have to relate this result either to a weakness in our definition of change or to molecular details that make disease-related but otherwise structural neutral mutations special.

Bibliography

1000 Genomes Project Consortium (2010). A map of human genome variation from population-scale sequencing. *Nature*, **467**(7319), 1061–1073.

Abagyan, R. A. and Batalov, S. (1997). Do aligned sequences share the same fold? *J Mol Biol*, **273**(1), 355–368.

Alber, T. and Matthews, B. W. (1987). Structure and thermal stability of phage T4 lysozyme. *Methods Enzymol*, **154**, 511–533.

Alber, T., Bell, J., Sun, D., Nicholson, H., Wozniak, J., Cook, S., and Matthews, B. (1988). Replacements of Pro86 in phage T4 lysozyme extend an alpha-helix but do not alter protein stability. *Science*, **239**(4840), 631–635.

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., and Walter, P. (2002). *Molecular biology of the cell*. Garland Science Taylor & Francis Group, 4 edition.

Alexander, P. A., He, Y., Chen, Y., Orban, J., and Bryan, P. N. (2009). A minimal sequence code for switching protein structure and function. *Proc Natl Acad Sci U S A*, **106**(50), 21149–21154.

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), 3389–3402.

Amberger, J., Bocchini, C. A., Scott, A. F., and Hamosh, A. (2009). McKusick's Online Mendelian Inheritance in Man (OMIM). *Nucleic Acids Res*, **37**(Database issue), D793–D796.

Andersen, C. A. F., Palmer, A. G., Brunak, S., and Rost, B. (2002). Continuum secondary structure captures protein flexibility. *Structure*, **10**(2), 175–184.

Anfinsen, C. B. (1973). Principles that govern the folding of protein chains. *Science*, **181**(4096), 223–230.

Baase, W. A., Eriksson, A. E., Zhang, X. J., Heinz, D. W., Sauer, U., Blaber, M., Baldwin, E. P., Wozniak, J. A., and Matthews, B. W. (1992). Dissection of protein structure and folding by directed mutagenesis. *Faraday Discuss*, (93), 173–181.

Bairoch, A., Apweiler, R., Wu, C. H., Barker, W. C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., Martin, M. J., Natale, D. A., O'Donovan, C., Redaschi, N., and Yeh, L.-S. L. (2005). The Universal Protein Resource (UniProt). *Nucleic Acids Res*, **33**(Database issue), D154–D159.

Bao, L. and Cui, Y. (2005). Prediction of the phenotypic effects of non-synonymous single nucleotide polymorphisms using structural and evolutionary information. *Bioinformatics*, **21**(10), 2185–2190.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000). The Protein Data Bank. *Nucleic Acids Res*, **28**(1), 235–242.

Betts, M. J. and Russell, R. B. (2003). *Amino Acid Properties and Consequences of Substitutions*, pages 289–316. John Wiley & Sons, Ltd.

Betz, S. F. (1993). Disulfide bonds and the stability of globular proteins. *Protein Sci*, **2**(10), 1551–1558.

Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M.-C., Estreicher, A., Gasteiger, E., Martin, M. J., Michoud, K., O'Donovan, C., Phan, I., Pilbout, S., and Schneider, M. (2003). The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res*, **31**(1), 365–370.

Bowie, J. U., Clarke, N. D., Pabo, C. O., and Sauer, R. T. (1990). Identification of protein folds: matching hydrophobicity patterns of sequence sets with solvent accessibility patterns of known structures. *Proteins*, **7**(3), 257–264.

Bowie, J. U., Lüthy, R., and Eisenberg, D. (1991). A method to identify protein sequences that fold into a known three-dimensional structure. *Science*, **253**(5016), 164–170.

Branden, C.-I. and Tooze, J. (1999). *Introduction to Protein Structure*. Garland Publishing, second edition.

Bromberg, Y. and Rost, B. (2007). SNAP: predict effect of non-synonymous polymorphisms on function. *Nucleic Acids Res*, **35**(11), 3823–3835.

Bromberg, Y. and Rost, B. (2008). Comprehensive in silico mutagenesis highlights functionally important residues in proteins. *Bioinformatics*, **24**(16), i207–i212.

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), 3059–3069.

Buckle, A. M., Cramer, P., and Fersht, A. R. (1996). Structural and energetic responses to cavity-creating mutations in hydrophobic cores: observation of a buried water molecule and the hydrophilic nature of such hydrophobic cavities. *Biochemistry*, **35**(14), 4298–4305.

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

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

Carpenter, E. P., Beis, K., Cameron, A. D., and Iwata, S. (2008). Overcoming the challenges of membrane protein crystallography. *Curr Opin Struct Biol*, **18**(5), 581–586.

Cerpa, R., Cohen, F. E., and Kuntz, I. D. (1996). Conformational switching in designed peptides: the helix/sheet transition. *Fold Des*, **1**(2), 91–101.

Chakravarti, A. (1998). It's raining SNPs, hallelujah? *Nat Genet*, **19**(3), 216–217.

Chan, H. S. and Dill, K. A. (1991). Polymer principles in protein structure and stability. *Annu Rev Biophys Biophys Chem*, **20**, 447–490.

Chasman, D. and Adams, R. M. (2001). Predicting the functional consequences of non-synonymous single nucleotide polymorphisms: structure-based assessment of amino acid variation. *J Mol Biol*, **307**(2), 683–706.

Chou, P. Y. and Fasman, G. D. (1978). Empirical predictions of protein conformation. *Annu Rev Biochem*, **47**, 251–276.

Chung, S. Y. and Subbiah, S. (1996). A structural explanation for the twilight zone of protein sequence homology. *Structure*, **4**(10), 1123–1127.

Collins, F. S., Guyer, M. S., and Charkravarti, A. (1997). Variations on a theme: cataloging human DNA sequence variation. *Science*, **278**(5343), 1580–1581.

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), 1229–1231.

Creighton, T. (1993). *Proteins: structures and molecular properties*. W.H. Freeman.

Crick, F. (1958). On protein synthesis. *Symp Soc Exp Biol*, **12**, 138–163.

Crick, F. (1970). Central dogma of molecular biology. *Nature*, **227**(5258), 561–563.

Dao-pin, S., Anderson, D. E., Baase, W. A., Dahlquist, F. W., and Matthews, B. W. (1991). Structural and thermodynamic consequences of burying a charged residue within the hydrophobic core of T4 lysozyme. *Biochemistry*, **30**(49), 11521–11529.

Dayhoff, M. O., Schwartz, R. M., and Orcutt, B. C. (1978). *Atlas of Protein Sequence and Structure*, chapter 5, pages 345–52.

De Filippis, V., Sander, C., and Vriend, G. (1994). Predicting local structural changes that result from point mutations. *Protein Eng*, **7**(10), 1203–1208.

Dill, K. A., Bromberg, S., Yue, K., Fiebig, K. M., Yee, D. P., Thomas, P. D., and Chan, H. S. (1995). Principles of protein folding–a perspective from simple exact models. *Protein Sci*, **4**(4), 561–602.

Doolittle, R. (1981). Similar amino acid sequences: chance or common ancestry? *Science*, **214**(4517), 149–159.

Dosztanyi, Z., Csizmok, V., Tompa, P., and Simon, I. (2005a). IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics*, **21**(16), 3433–3434.

Dosztanyi, Z., Csizmok, V., Tompa, P., and Simon, I. (2005b). The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. *J Mol Biol*, **347**(4), 827–839.

Dunker, A. K., Obradovic, Z., Romero, P., Garner, E. C., and Brown, C. J. (2000). Intrinsic protein disorder in complete genomes. *Genome Inform Ser Workshop Genome Inform*, **11**, 161–171.

Dunker, A. K., Lawson, J. D., Brown, C. J., Williams, R. M., Romero, P., Oh, J. S., Oldfield, C. J., Campen, A. M., Ratliff, C. M., Hipps, K. W., Ausio, J., Nissen, M. S., Reeves, R., Kang, C., Kissinger, C. R., Bailey, R. W., Griswold, M. D., Chiu, W., Garner, E. C., and Obradovic, Z. (2001). Intrinsically disordered protein. *J Mol Graph Model*, **19**(1), 26–59.

Dunker, A. K., Brown, C. J., Lawson, J. D., Iakoucheva, L. M., and Obradović, Z. (2002). Intrinsic disorder and protein function. *Biochemistry*, **41**(21), 6573–6582.

Dyson, H. J. and Wright, P. E. (2005). Intrinsically unstructured proteins and their functions. *Nat Rev Mol Cell Biol*, **6**(3), 197–208.

Eriksson, A. E., Baase, W. A., Zhang, X. J., Heinz, D. W., Blaber, M., Baldwin, E. P., and Matthews, B. W. (1992). Response of a protein structure to cavity-creating mutations and its relation to the hydrophobic effect. *Science*, **255**(5041), 178–183.

Eyal, E., Najmanovich, R., Sobolev, V., and Edelman, M. (2001). MutaProt: a web interface for structural analysis of point mutations. *Bioinformatics*, **17**(4), 381–382.

Eyal, E., Najmanovich, R., Edelman, M., and Sobolev, V. (2003). Protein side-chain rearrangement in regions of point mutations. *Proteins*, **50**(2), 272–282.

Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R., and Lin, C.-J. (2008). LIBLINEAR: A Library for Large Linear Classification. *Journal of Machine Learning Research*, **9**, 1871–1874.

Fernald, G. H., Capriotti, E., Daneshjou, R., Karczewski, K. J., and Altman, R. B. (2011). Bioinformatics challenges for personalized medicine. *Bioinformatics*, **27**(13), 1741–1748.

Ferraroni, M., Rypniewski, W., Wilson, K. S., Viezzoli, M. S., Banci, L., Bertini, I., and Mangani, S. (1999). The crystal structure of the monomeric human SOD mutant F50E/G51E/E133Q at atomic resolution. The enzyme mechanism revisited. *J Mol Biol*, **288**(3), 413–426.

Ferrer-Costa, C., Orozco, M., and de la Cruz, X. (2002). Characterization of disease-associated single amino acid polymorphisms in terms of sequence and structure properties. *J Mol Biol*, **315**(4), 771–786.

Finn, R. D., Mistry, J., Tate, J., Coggill, P., Heger, A., Pollington, J. E., Gavin, O. L., Gunasekaran, P., Ceric, G., Forslund, K., Holm, L., Sonnhammer, E. L. L., Eddy, S. R., and Bateman, A. (2010). The Pfam protein families database. *Nucleic Acids Res*, **38**(Database issue), D211–D222.

Finn, R. D., Clements, J., and Eddy, S. R. (2011). HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res*, **39**(Web Server issue), W29–W37.

Fliess, A., Motro, B., and Unger, R. (2002). Swaps in protein sequences. *Proteins*, **48**(2), 377–387.

Franzosa, E. A. and Xia, Y. (2009). Structural determinants of protein evolution are context-sensitive at the residue level. *Mol Biol Evol*, **26**(10), 2387–2395.

Galea, C., Bowman, P., and Kriwacki, R. W. (2005). Disruption of an intermonomer salt bridge in the p53 tetramerization domain results in an increased propensity to form amyloid fibrils. *Protein Sci*, **14**(12), 2993–3003.

Garcia-Seisdedos, H., Ibarra-Molero, B., and Sanchez-Ruiz, J. M. (2012). How many ionizable groups can sit on a protein hydrophobic core? *Proteins*, **80**(1), 1–7.

Gong, S. and Blundell, T. L. (2010). Structural and functional restraints on the occurrence of single amino acid variations in human proteins. *PLoS One*, **5**(2), e9186.

Goodman, M., Braunitzer, G., Stangl, A., and Schrank, B. (1983). Evidence on human origins from haemoglobins of African apes. *Nature*, **303**(5917), 546–548.

Gooptu, B. and Lomas, D. A. (2009). Conformational pathology of the serpins: themes, variations, and therapeutic strategies. *Annu Rev Biochem*, **78**, 147–176.

Gray, T. M., Arnoys, E. J., Blankespoor, S., Born, T., Jagar, R., Everman, R., Plowman, D., Stair, A., and Zhang, D. (1996). Destabilizing effect of proline substitutions in two helical regions of T4 lysozyme: leucine 66 to proline and leucine 91 to proline. *Protein Sci*, **5**(4), 742–751.

Guy, H. R. (1985). Amino acid side-chain partition energies and distribution of residues in soluble proteins. *Biophys J*, **47**(1), 61–70.

Guyon, I. and Elisseeff, A. (2003). An introduction to variable and feature selection. *J. Mach. Learn. Res.*, **3**, 1157–1182.

Hecht, M. H., Nelson, H. C., and Sauer, R. T. (1983). Mutations in lambda repressor's amino-terminal domain: implications for protein stability and DNA binding. *Proc Natl Acad Sci U S A*, **80**(9), 2676–2680.

Hu, Y., Liu, Y., Jung, J., Dunker, A. K., and Wang, Y. (2011). Changes in predicted protein disorder tendency may contribute to disease risk. *BMC Genomics*, **12 Suppl 5**, S2.

International HapMap Consortium (2003). The International HapMap Project. *Nature*, **426**(6968), 789–796.

Isom, D. G., Castaneda, C. A., Cannon, B. R., Velu, P. D., and Garcia-Moreno E, B. (2010). Charges in the hydrophobic interior of proteins. *Proc Natl Acad Sci U S A*, **107**(37), 16096–16100.

Janin, J. and Chothia, C. (1990). The structure of protein-protein recognition sites. *J Biol Chem*, **265**(27), 16027–16030.

Janin, J., Miller, S., and Chothia, C. (1988). Surface, subunit interfaces and interior of oligomeric proteins. *J Mol Biol*, **204**(1), 155–164.

Joerger, A. C., Ang, H. C., and Fersht, A. R. (2006). Structural basis for understanding oncogenic p53 mutations and designing rescue drugs. *Proc Natl Acad Sci U S A*, **103**(41), 15056–15061.

Jones, S. and Thornton, J. M. (1996). Principles of protein-protein interactions. *Proc Natl Acad Sci U S A*, **93**(1), 13–20.

Kabsch, W. (1976). A solution for the best rotation to relate two sets of vectors. *Acta Crystallographica Section A*, **32**(5), 922–923.

Kabsch, W. (1978). A discussion of the solution for the best rotation to relate two sets of vectors. *Acta Crystallographica Section A*, **34**(5), 827–828.

Kabsch, W. and Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**(12), 2577–2637.

Kabsch, W. and Sander, C. (1984). On the use of sequence homologies to predict protein structure: identical pentapeptides can have completely different conformations. *Proc Natl Acad Sci U S A*, **81**(4), 1075–1078.

Kamtekar, S., Schiffer, J. M., Xiong, H., Babik, J. M., and Hecht, M. H. (1993). Protein design by binary patterning of polar and nonpolar amino acids. *Science*, **262**(5140), 1680–1685.

Kawabata, T., Ota, M., and Nishikawa, K. (1999). The Protein Mutant Database. *Nucleic Acids Res*, **27**(1), 355–357.

Kendrew, J. C., Bodo, G., Dintzis, H. M., Parrish, R. G., Wyckoff, H., and Phillips, D. C. (1958). A Three-Dimensional Model of the Myoglobin Molecule Obtained by X-Ray Analysis. *Nature*, **181**(4610), 662–666.

Kimura, M. (1983). *The neutral theory of molecular evolution*. Cambridge University Press.

Kircher, M. and Kelso, J. (2010). High-throughput DNA sequencing–concepts and limitations. *Bioessays*, **32**(6), 524–536.

Korndörfer, I., Steipe, B., Huber, R., Tomschy, A., and Jaenicke, R. (1995). The crystal structure of holo-glyceraldehyde-3-phosphate dehydrogenase from the hyperthermophilic bacterium Thermotoga maritima at 2.5 A resolution. *J Mol Biol*, **246**(4), 511–521.

Krishnan, V. G. and Westhead, D. R. (2003). A comparative study of machine-learning methods to predict the effects of single nucleotide polymorphisms on protein function. *Bioinformatics*, **19**(17), 2199–2209.

Kumar, M. D. S., Bava, K. A., Gromiha, M. M., Prabakaran, P., Kitajima, K., Uedaira, H., and Sarai, A. (2006). ProTherm and ProNIT: thermodynamic databases for proteins and protein-nucleic acid interactions. *Nucleic Acids Res*, **34**(Database issue), D204–D206.

Kyte, J. and Doolittle, R. F. (1982). A simple method for displaying the hydropathic character of a protein. *J Mol Biol*, **157**(1), 105–132.

Lei, K. J., Shelly, L. L., Pan, C. J., Sidbury, J. B., and Chou, J. Y. (1993). Mutations in the glucose-6-phosphatase gene that cause glycogen storage disease type 1a. *Science*, **262**(5133), 580–583.

Lesk, A. M. and Chothia, C. (1980). How different amino acid sequences determine similar protein structures: the structure and evolutionary dynamics of the globins. *J Mol Biol*, **136**(3), 225–270.

Li, W. and Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, **22**(13), 1658–1659.

Liu, J., Tan, H., and Rost, B. (2002). Loopy proteins appear conserved in evolution. *J Mol Biol*, **322**(1), 53–64.

Liu, R., Baase, W. A., and Matthews, B. W. (2000). The introduction of strain and its effects on the structure and stability of T4 lysozyme. *J Mol Biol*, **295**(1), 127–145.

Matthews, B. W. (1993). Structural and genetic analysis of protein stability. *Annu Rev Biochem*, **62**, 139–160.

Mattick, J. S. (2009). Deconstructing the dogma: a new view of the evolution and genetic programming of complex organisms. *Ann N Y Acad Sci*, **1178**, 29–46.

McGill, R., Tukey, J. W., and Larsen, W. A. (1978). Variations of Box Plots. *The American Statistician*, **32**(1), pp. 12–16.

McLachlan, A. D. (1982). Rapid comparison of protein structures. *Acta Crystallographica Section A*, **38**(6), 871–873.

Mendez, M., Sorkin, L., Rossetti, M. V., Astrin, K. H., del C Batlle, A. M., Parera, V. E., Aizencang, G., and Desnick, R. J. (1998). Familial porphyria cutanea tarda: characterization of seven novel uroporphyrinogen decarboxylase mutations and frequency of common hemochromatosis alleles. *Am J Hum Genet*, **63**(5), 1363–1375.

Mika, S. and Rost, B. (2003). UniqueProt: Creating representative protein sequence sets. *Nucleic Acids Res*, **31**(13), 3789–3791.

Murzin, A. G., Brenner, S. E., Hubbard, T., and Chothia, C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol*, **247**(4), 536–540.

Mészáros, B., Tompa, P., Simon, I., and Dosztányi, Z. (2007). Molecular principles of the interactions of disordered proteins. *J Mol Biol*, **372**(2), 549–561.

Ng, P. C. and Henikoff, S. (2001). Predicting deleterious amino acid substitutions. *Genome Res*, **11**(5), 863–874.

Ng, P. C. and Henikoff, S. (2003). SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res*, **31**(13), 3812–3814.

Obradovic, Z., Peng, K., Vucetic, S., Radivojac, P., and Dunker, A. K. (2005). Exploiting heterogeneous sequence properties improves prediction of protein disorder. *Proteins*, **61 Suppl 7**, 176–182.

Ofran, Y. and Rost, B. (2003). Analysing six types of protein-protein interfaces. *J Mol Biol*, **325**(2), 377–387.

Ofran, Y. and Rost, B. (2007a). ISIS: interaction sites identified from sequence. *Bioinformatics*, **23**(2), e13–e16.

Ofran, Y. and Rost, B. (2007b). Protein-protein interaction hotspots carved into sequences. *PLoS Comput Biol*, **3**(7), e119.

Ofran, Y., Mysore, V., and Rost, B. (2007). Prediction of DNA-binding residues from sequence. *Bioinformatics*, **23**(13), i347–i353.

Ozbabacan, S. E. A., Engin, H. B., Gursoy, A., and Keskin, O. (2011). Transient protein-protein interactions. *Protein Eng Des Sel*, **24**(9), 635–648.

Pakula, A. A., Young, V. B., and Sauer, R. T. (1986). Bacteriophage lambda cro mutations: effects on activity and intracellular degradation. *Proc Natl Acad Sci U S A*, **83**(23), 8829–8833.

Peng, K., Radivojac, P., Vucetic, S., Dunker, A. K., and Obradovic, Z. (2006). Length-dependent prediction of protein intrinsic disorder. *BMC Bioinformatics*, **7**, 208.

Perutz, M., Kendrew, J., and Watson, H. (1965). Structure and function of haemoglobin: II. Some relations between polypeptide chain configuration and amino acid sequence. *Journal of Molecular Biology*, **13**(3), 669 – 678.

Perutz, M. F. and Raidt, H. (1975). Stereochemical basis of heat stability in bacterial ferredoxins and in haemoglobin A2. *Nature*, **255**(5505), 256–259.

Perutz, M. F., Rossmann, M. G., Cullis, A. F., Muirhead, H., Will, G., and North, A. C. T. (1960). Structure of Haemoglobin: A Three-Dimensional Fourier Synthesis at 5.5-[angst]. Resolution, Obtained by X-Ray Analysis. *Nature*, **185**(4711), 416–422.

Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., and Ferrin, T. E. (2004). UCSF Chimera–a visualization system for exploratory research and analysis. *J Comput Chem*, **25**(13), 1605–1612.

Ponder, J. W. and Richards, F. M. (1987). Tertiary templates for proteins: Use of packing criteria in the enumeration of allowed sequences for different structural classes. *Journal of Molecular Biology*, **193**(4), 775 – 791.

Pruitt, K. D., Tatusova, T., and Maglott, D. R. (2005). NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res*, **33**(Database issue), D501–D504.

Ratovitski, T., Corson, L. B., Strain, J., Wong, P., Cleveland, D. W., Culotta, V. C., and Borchelt, D. R. (1999). Variation in the biochemical/biophysical properties of mutant superoxide dismutase 1 enzymes and the rate of disease progression in familial amyotrophic lateral sclerosis kindreds. *Hum Mol Genet*, **8**(8), 1451–1460.

Reichert, E. T. and Brown, A. P. (1909). *The differentiation and specificity of corresponding proteins and other vital substances in relation to biological classification and organic evolution: the crystallography of hemoglobins*. Carnegie Institution of Washington.

Richardson, J. S. and Richardson, D. C. (1988). Amino acid preferences for specific locations at the ends of alpha helices. *Science*, **240**(4859), 1648–1652.

Roessler, C. G., Hall, B. M., Anderson, W. J., Ingram, W. M., Roberts, S. A., Montfort, W. R., and Cordes, M. H. J. (2008). Transitive homology-guided structural studies lead to discovery of Cro proteins with 40% sequence identity but different folds. *Proc Natl Acad Sci U S A*, **105**(7), 2343–2348.

Romero, P., Obradovic, Z., Li, X., Garner, E. C., Brown, C. J., and Dunker, A. K. (2001). Sequence complexity of disordered protein. *Proteins*, **42**(1), 38–48.

Rose, G. D., Gierasch, L. M., and Smith, J. A. (1985). Turns in peptides and proteins. *Adv Protein Chem*, **37**, 1–109.

Rost, B. (1996). PHD: predicting one-dimensional protein structure by profile-based neural networks. *Methods Enzymol*, **266**, 525–539.

Rost, B. (1997). Protein structures sustain evolutionary drift. *Fold Des*, **2**(3), S19–S24.

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

Rost, B. (2005). How to Use Protein 1- D Structure Predicted by PROFphd. In J. M. Walker, editor, *The Proteomics Protocols Handbook*, pages 875–901. Humana Press. 10.1385/1-59259-890-0:875.

Rost, B. and Sander, C. (1993). Prediction of protein secondary structure at better than 70accuracy. *J Mol Biol*, **232**(2), 584–599.

Rost, B. and Sander, C. (1994). Combining evolutionary information and neural networks to predict protein secondary structure. *Proteins*, **19**(1), 55–72.

Rost, B., Sander, C., and Schneider, R. (1994). Redefining the goals of protein secondary structure prediction. *J Mol Biol*, **235**(1), 13–26.

Rost, B., Schneider, R., and Sander, C. (1997). Protein fold recognition by prediction-based threading. *J Mol Biol*, **270**(3), 471–480.

Rost, B., Yachdav, G., and Liu, J. (2004). The PredictProtein server. *Nucleic Acids Research*, **32**(Web-Server-Issue), 321–326.

Sandberg, W. S. and Terwilliger, T. C. (1989). Influence of interior packing and hydrophobicity on the stability of a protein. *Science*, **245**(4913), 54–57.

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

Saunders, C. T. and Baker, D. (2002). Evaluation of structural and evolutionary contributions to deleterious mutation prediction. *J Mol Biol*, **322**(4), 891–901.

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

Schaefer, C., Schlessinger, A., and Rost, B. (2010). Protein secondary structure appears to be robust under in silico evolution while protein disorder appears not to be. *Bioinformatics*, **26**(5), 625–631.

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

Schaefer, C., Meier, A., Rost, B., and Bromberg, Y. (2012b). SNPdbe: constructing an nsSNP functional impacts database. *Bioinformatics*, **28**(4), 601–602.

Schimmel, P. R. and Flory, P. J. (1968). Conformational energies and configurational statistics of copolypeptides containing L-proline. *J Mol Biol*, **34**(1), 105–120.

Schlessinger, A., Yachdav, G., and Rost, B. (2006). PROFbval: predict flexible and rigid residues in proteins. *Bioinformatics*, **22**(7), 891–893.

Schlessinger, A., Liu, J., and Rost, B. (2007a). Natively unstructured loops differ from other loops. *PLoS Comput Biol*, **3**(7), e140.

Schlessinger, A., Punta, M., and Rost, B. (2007b). Natively unstructured regions in proteins identified from contact predictions. *Bioinformatics*, **23**(18), 2376–2384.

Schlessinger, A., Punta, M., Yachdav, G., Kajan, L., and Rost, B. (2009). Improved disorder prediction by combination of orthogonal approaches. *PLoS One*, **4**(2), e4433.

Schlessinger, A., Schaefer, C., Vicedo, E., Schmidberger, M., Punta, M., and Rost, B. (2011). Protein disorder–a breakthrough invention of evolution? *Curr Opin Struct Biol*, **21**(3), 412–418.

Shakhnovich, E. I. and Gutin, A. M. (1991). Influence of point mutations on protein structure: probability of a neutral mutation. *J Theor Biol*, **149**(4), 537–546.

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), 308–311.

Shieh, J.-J., Terzioglu, M., Hiraiwa, H., Marsh, J., Pan, C.-J., Chen, L.-Y., and Chou, J. Y. (2002). The molecular basis of glycogen storage disease type 1a: structure and function analysis of mutations in glucose-6-phosphatase. *J Biol Chem*, **277**(7), 5047–5053.

Shirley, B. A., Stanssens, P., Hahn, U., and Pace, C. N. (1992). Contribution of hydrogen bonding to the conformational stability of ribonuclease T1. *Biochemistry*, **31**(3), 725–732.

Shortle, D. and Lin, B. (1985). Genetic analysis of staphylococcal nuclease: identification of three intragenic "global" suppressors of nuclease-minus mutations. *Genetics*, **110**(4), 539–555.

Shortle, D., Stites, W. E., and Meeker, A. K. (1990). Contributions of the large hydrophobic amino acids to the stability of staphylococcal nuclease. *Biochemistry*, **29**(35), 8033–8041.

Smialowski, P., Frishman, D., and Kramer, S. (2010). Pitfalls of supervised feature selection. *Bioinformatics*, **26**(3), 440–443.

Smith, M., Kunin, V., Goldovsky, L., Enright, A. J., and Ouzounis, C. A. (2005). MagicMatch–cross-referencing sequence identifiers across databases. *Bioinformatics*, **21**(16), 3429–3430.

Stitziel, N. O., Tseng, Y. Y., Pervouchine, D., Goddeau, D., Kasif, S., and Liang, J. (2003). Structural location of disease-associated single-nucleotide polymorphisms. *J Mol Biol*, **327**(5), 1021–1030.

Sunyaev, S., Ramensky, V., and Bork, P. (2000). Towards a structural basis of human non-synonymous single nucleotide polymorphisms. *Trends Genet*, **16**(5), 198–200.

Sunyaev, S., Ramensky, V., Koch, I., Lathe, 3rd, W., Kondrashov, A. S., and Bork, P. (2001). Prediction of deleterious human alleles. *Hum Mol Genet*, **10**(6), 591–597.

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), 387–394.

Talavera, D., Taylor, M. S., and Thornton, J. M. (2010). The (non)malignancy of cancerous amino acidic substitutions. *Proteins*, **78**(3), 518–529.

Tanford, C. (1978). The hydrophobic effect and the organization of living matter. *Science*, **200**(4345), 1012–1018.

Taylor, T. J. and Vaisman, I. I. (2010). Discrimination of thermophilic and mesophilic proteins. *BMC Struct Biol*, **10 Suppl 1**, S5.

Tukey, J. (1977). *Exploratory data analysis*. Addison Wesley, 1st edition.

Uversky, V. N., Oldfield, C. J., and Dunker, A. K. (2005). Showing your ID: intrinsic disorder as an ID for recognition, regulation and cell signaling. *J Mol Recognit*, **18**(5), 343–384.

Vacic, V. and Iakoucheva, L. M. (2012). Disease mutations in disordered regions–exception to the rule? *Mol Biosyst*, **8**(1), 27–32.

Vetriani, C., Maeder, D. L., Tolliday, N., Yip, K. S., Stillman, T. J., Britton, K. L., Rice, D. W., Klump, H. H., and Robb, F. T. (1998). Protein thermostability above 100 degreesC: a key role for ionic interactions. *Proc Natl Acad Sci U S A*, **95**(21), 12300–12305.

Vinogradov, S. N., Hoogewijs, D., Bailly, X., Mizuguchi, K., Dewilde, S., Moens, L., and Vanfleteren, J. R. (2007). A model of globin evolution. *Gene*, **398**(1-2), 132–142.

Vucetic, S., Obradovic, Z., Vacic, V., Radivojac, P., Peng, K., Iakoucheva, L. M., Cortese, M. S., Lawson, J. D., Brown, C. J., Sikes, J. G., Newton, C. D., and Dunker, A. K. (2005). DisProt: a database of protein disorder. *Bioinformatics*, **21**(1), 137–140.

Vucetic, S., Xie, H., Iakoucheva, L. M., Oldfield, C. J., Dunker, A. K., Obradovic, Z., and Uversky, V. N. (2007). Functional anthology of intrinsic disorder. 2. Cellular components, domains, technical terms, developmental processes, and coding sequence diversities correlated with long disordered regions. *J Proteome Res*, **6**(5), 1899–1916.

Wang, K., Li, M., and Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res*, **38**(16), e164.

Wang, Z. and Moult, J. (2001). SNPs, protein structure, and disease. *Hum Mutat*, **17**(4), 263–270.

Wang, Z. and Moult, J. (2003). Three-dimensional structural location and molecular functional effects of missense SNPs in the T cell receptor Vbeta domain. *Proteins*, **53**(3), 748–757.

Weaver, L. H., Gray, T. M., Grutter, M. G., Anderson, D. E., Wozniak, J. A., Dahlquist, F. W., and Matthews, B. W. (1989). High-resolution structure of the temperature-sensitive mutant of phage lysozyme, Arg 96—-His. *Biochemistry*, **28**(9), 3793–3797.

Xu, J., Baase, W. A., Baldwin, E., and Matthews, B. W. (1998). The response of T4 lysozyme to large-to-small substitutions within the core and its relation to the hydrophobic effect. *Protein Sci*, **7**(1), 158–177.

Ye, Y. and Godzik, A. (2003). Flexible structure alignment by chaining aligned fragment pairs allowing twists. *Bioinformatics*, **19 Suppl 2**, ii246–ii255.

Ye, Z.-Q., Zhao, S.-Q., Gao, G., Liu, X.-Q., Langlois, R. E., Lu, H., and Wei, L. (2007). Finding new structural and sequence attributes to predict possible disease association of single amino acid polymorphism (SAP). *Bioinformatics*, **23**(12), 1444–1450.

Yip, K. S., Stillman, T. J., Britton, K. L., Artymiuk, P. J., Baker, P. J., Sedelnikova, S. E., Engel, P. C., Pasquo, A., Chiaraluce, R., and Consalvi, V. (1995). The structure of Pyrococcus furiosus glutamate dehydrogenase reveals a key role for ion-pair networks in maintaining enzyme stability at extreme temperatures. *Structure*, **3**(11), 1147–1158.

Yip, Y. L., Famiglietti, M., Gos, A., Duek, P. D., David, F. P. A., Gateau, A., and Bairoch, A. (2008). Annotating single amino acid polymorphisms in the UniProt/Swiss-Prot knowledgebase. *Hum Mutat*, **29**(3), 361–366.

Zamyatnin, A. A. (1972). Protein volume in solution. *Prog Biophys Mol Biol*, **24**, 107–123.

Zuckerkandl, E. and Pauling, L. (1965). Evolutionary divergence and convergence in proteins. In V. Bryson and H. J. Vogel, editors, *Evolving Genes and Proteins*, pages 97–166. Academic Press.

Acknowledgements

The dissertation at hand constitutes the aggregation of over three years filled with research and more. I wish to thank those who made this possible.

First and foremost I want to thank Burkhard Rost for his support in many ways. For me, it was an honor to have been accepted into his group and to get the opportunity to learn from him. Beyond that, I am very grateful to Burkhard for the great time I had in one of the most amazing cities.

Just as well, I thank Ulrich Mansmann, professor and head of the epidemiological institute at Klinikum Grosshadern Munich, for his guidance and helpful discussions.

Also, I am indebted to my fellow colleagues Marco Punta, Yana Bromberg, Laszlo Kajan, Marc Offman, Andrea Schafferhans, Shaila Roessle, Edda Kloppmann, Markus Schmidberger and Avner Schlessinger for their collaborations, their work and for having shared their knowledge with me.

In this respect, I am obliged to Alice Meier, Yannick Mahlich and Dominik Achten whose significant work as undergraduate students helped to shape this dissertation. Special thanks go to Marlena Drabik and Timothy Karl. Furthermore, I would like to thank the whole lab for inspiration and time spent together beyond science.

Thanks also go to the anonymous reviewers who shaped my publications with their fruitful comments and concerns. Just as well, I wish to thank the contributors to and the teams from PDB, dbSNP, UniProt, PMD, OMIM, DisProt and RefSeq. This work would not have been possible without their public available data.

Last but not least, thanks to my parents Ulrike and Bernd who supported me in so many ways.

Appendix

The manuscripts of the following peer-reviewed publications have been appended:

- **Christian Schaefer**, Avner Schlessinger, Burkhard Rost (2010). **Protein secondary structure appears to be robust under in silico evolution while protein disorder appears not to be.** Bioinformatics, 26(5):625-631.
- **Christian Schaefer**, Alice Meier, Burkhard Rost, Yana Bromberg (2012). **SNPdbe: Constructing an nsSNP functional impacts database.** Bioinformatics, 28(4):601-2.
- **Christian Schaefer**, Burkhard Rost (2012). **Predict impact of single amino acid change upon protein structure.** BMC Genomics, 13(Suppl 4):S4.
- **Christian Schaefer**, Yana Bromberg, Dominik Achten, Burkhard Rost (2012). **Disease-related mutations predicted to impact protein function.** BMC Genomics, 13(Suppl 4):S11.

Summaries of the publications and my individual contributions are as follows.

Protein secondary structure appears to be robust under in silico evolution while protein disorder appears not to be

The mutation of single amino acids in proteins often impacts protein function and structure. Those exceptional mutations that have no negative effect sustain evolutionary pressure.

In this publication, we studied a particular aspect of robustness with respect to mutations, namely regular protein secondary structure (helices and sheets) and natively unstructured or intrinsically disordered regions (often observed in non-regular secondary structure). Is the formation of regular secondary structure an intrinsic feature of amino acid sequences, or is it a feature that is easily lost upon mutation and is maintained by evolution against the odds? Similarly, is disorder an intrinsic sequence feature or is it difficult to maintain?

To tackle these questions, we in-silico mutated native protein sequences gradually into random sequence-like ensembles and monitored the change in predicted secondary structure through PROFsec (Rost and Sander, 1993) and disorder through IUPred (Dosztanyi *et al.*, 2005a,b), MD (Schlessinger *et al.*, 2009) and VSL2 (Obradovic *et al.*, 2005; Peng *et al.*, 2006).

We established that by our coarse-grained measures for change, predictions and observations were indeed similar for the native sequences. The strings of secondary structure (three states) and disorder (two states) began to differ from the native one at a rate roughly linearly proportional to the change in sequence. Surprisingly, neither the content in regular secondary structure nor the length distribution of helices and strands changed substantially; instead, helices and strands were lost and created at similar rates. Regions with long disorder (>30 consecutive residues) behaved very differently: they just disappeared during our in silico mutations.

Our findings suggest that the ability to form regular secondary structure is an intrinsic feature of amino acid sequences from well-ordered proteins, while the ability to form disordered regions is significantly less an intrinsic feature of proteins with disordered regions. Put differently: helices and strands are easy to maintain by evolution, whereas disordered regions are difficult to maintain. Mutations that are neutral with respect to disorder are therefore extremely unlikely.

The study design and methodology were conceived by myself and Burkhard Rost. I carried out necessary background research. The programming was performed by me with the help of Avner Schlessinger. All calculations were done by myself with the help of Burkhard Rost. The resulting data were analyzed and interpreted by myself and Burkhard Rost. The manuscript was drafted by myself, Avner Schlessinger and Burkhard Rost.

SNPdbe: Constructing an nsSNP functional impacts database

Many existing databases annotate experimentally characterized single nucleotide polymorphisms (SNPs; most prominently dbSNP (Sherry *et al.*, 2001)). Each non-synonymous SNP (nsSNP) changes one amino acid in the gene product. This change can either affect protein function or be neutral in that respect. Most polymorphisms lack experimental annotation of their functional impact.

In this publication, we introduced SNPdbe – SNP database of effects, with predictions of computationally annotated functional impacts of SNPs. Database entries were derived from nsSNPs in dbSNP, and variants reported in Swiss-Prot (Bairoch *et al.*, 2005), SwissVar (Yip *et al.*, 2008), 1000 Genomes Project (1000 Genomes Project Consortium, 2010) and PMD (Kawabata *et al.*, 1999). nsSNPs come from more than 2600 organisms; "human" being the most prevalent. The impact of each nsSNP on protein function was predicted using the SNAP (Bromberg and Rost, 2007) and SIFT (Ng and Henikoff, 2003, 2001) algorithms and augmented with experimentally derived function/structure information and disease associations from PMD, SwissVar and OMIM (Amberger *et al.*, 2009). SNPdbe is consistently updated and easily augmented with new sources of information.

The database is available as a MySQL dump and via a web front-end that allows searching using any combination of organism names, sequences and mutation IDs.

The methodology was conceived by myself and Yana Bromberg. I carried out necessary background research. The programming, data collection and database development were performed by myself. The web frontend was programmed by Alice Meier under my supervision. The manuscript was drafted by myself, Burkhard Rost and Yana Bromberg.

Predict impact of single amino acid change upon protein structure

Changing a single amino acid in a protein potentially affects protein structure, function and phenotype such as disease. In this publication, we proposed a direct method that predicts the impact of single amino changes upon local protein structure. We compiled a data collection out of the PDB (Berman *et al.*, 2000) consisting of structurally superimposed protein fragment pairs of five consecutive residues. Each such pentamer pair had the following properties: Both peptides shared identical flanking regions in sequence but had one mismatch position in their center. We inferred the structural effect imposed by the central mismatch by measuring the root mean square displacement (RMSD) between two fragments. We defined pairs having a RMSD < 0.2Å as structurally neutral and > 0.4Å as structurally non-neutral. We applied logistic regression (Fan *et al.*, 2008) to machine-learn the effects of mismatches on local structure.

We established a seemingly rather high overall performance (AUC>0.79, two-state accuracy 72.6%). Despite this success, our method largely failed to discriminate between the effects of changes upon stability and function. Nonetheless, mutants for which our method predicted a change of structure were also enriched in terms of disrupting stability and function.

Our definition for structural change enabled the application of machine-learning to distinguish structural neutral from non-neutral. But we failed in the next step, namely to use predicted local structural changes to infer the impact of a mutation upon protein stability and/or function. This might be due to our particular definition of structural change.

The study design and methodology were conceived by myself and Burkhard Rost. I carried out necessary background research. The programming was performed by me. All calculations were done by myself with the help of Burkhard Rost. The resulting data were analyzed and interpreted by myself and Burkhard Rost. The manuscript was drafted by myself and Burkhard Rost.

Disease-related mutations predicted to impact protein function

Genomic point mutations that alter the protein sequence (non-synonymous single nucleotide polymorphisms, nsSNPs) are of distinct interest: they could influence the phenotype by, e.g., causing disease. However, few are annotated with function and even fewer map to diseases. What are the underlying molecular mechanisms that make one mutation functional neutral, deleterious or even disease causing?

In this publication, we studied the relationship between functional change upon point mutation and disease. We used disease-annotated variants from SwissVar (Yip *et al.*, 2008), OMIM (Amberger *et al.*, 2009) and PMD (Kawabata *et al.*, 1999) and variants not linked to disease and predicted their functional impact using SNAP (Bromberg and Rost, 2007) and SIFT (Ng and Henikoff, 2003, 2001) algorithms.

Mutations predicted to effect protein function were more abundant in disease-causing variants than mutations predicted to be neutral. Even more surprising, we found that the predictions of mutations that effect function were much stronger for nsSNPs annotated to cause disease than for other data sets annotating functional change. Our findings suggest that for the majority of disease mutants loss-of-function is an essential disease-causing factor. Conversely, we confirmed that not all mutations predicted to change function are related to disease and that some mutations predicted to be neutral are annotated as related to disease. Hence, a clear one-to-one relation between function and disease remains elusive.

The study design and methodology were conceived by myself and Burkhard Rost. I carried out necessary background research. The programming and data collection were mainly performed by me with the help of Dominik Achten. All calculations were done mainly by myself with the help of Burkhard Rost. The resulting data were analyzed and interpreted by myself, Yana Bromberg and Burkhard Rost. The manuscript was drafted by myself, Yana Bromberg and Burkhard Rost.

BIOINFORMATICS ORIGINAL PAPER

Vol. 26 no. 5 2010, pages 625–631
doi:10.1093/bioinformatics/btq012

Structural bioinformatics

Advance Access publication January 16, 2010

Protein secondary structure appears to be robust under *in silico* evolution while protein disorder appears not to be

Christian Schaefer[1,2,*], Avner Schlessinger[3] and Burkhard Rost[1,2,4]

[1]Department of Biochemistry and Molecular Biophysics, Center for Computational Biology and Bioinformatics (C2B2), Columbia University, 1130 St Nicholas Ave. Rm. 802, New York, NY 10032, USA, [2]Department of Computer Science, Institute of Advanced Studies (IAS), NorthEast Structural Genomics Consortium (NESG), TUM Bioinformatics, TUM Munich, Boltzmannstr. 3, 85748 Garching, Germany, [3]Department of Bioengineering and Therapeutic Sciences, University of California at San Francisco, 1700, 4th Street, San Francisco, CA 94158 and [4]NorthEast Structural Genomics Consortium (NESG) and New York Consortium on Membrane Protein Structure (NYCOMPS), 1130 St. Nicholas Ave. Rm. 802, New York, NY 10032, USA

Associate Editor: Alfonso Valencia

ABSTRACT

Motivation: The mutation of amino acids often impacts protein function and structure. Mutations without negative effect sustain evolutionary pressure. We study a particular aspect of structural robustness with respect to mutations: regular protein secondary structure and natively unstructured (intrinsically disordered) regions. Is the formation of regular secondary structure an intrinsic feature of amino acid sequences, or is it a feature that is lost upon mutation and is maintained by evolution against the odds? Similarly, is disorder an intrinsic sequence feature or is it difficult to maintain? To tackle these questions, we *in silico* mutated native protein sequences into random sequence-like ensembles and monitored the change in predicted secondary structure and disorder.

Results: We established that by our coarse-grained measures for change, predictions and observations were similar, suggesting that our results were not biased by prediction mistakes. Changes in secondary structure and disorder predictions were linearly proportional to the change in sequence. Surprisingly, neither the content nor the length distribution for the predicted secondary structure changed substantially. Regions with long disorder behaved differently in that significantly fewer such regions were predicted after a few mutation steps. Our findings suggest that the formation of regular secondary structure is an intrinsic feature of random amino acid sequences, while the formation of long-disordered regions is not an intrinsic feature of proteins with disordered regions. Put differently, helices and strands appear to be maintained easily by evolution, whereas maintaining disordered regions appears difficult. Neutral mutations with respect to disorder are therefore very unlikely.

Contact: schaefer@rostlab.org

Supplementary Information: Supplementary data are available at *Bioinformatics* online.

Received on August 4, 2009; revised on December 18, 2009; accepted on January 9, 2010

*To whom correspondence should be addressed.

1 INTRODUCTION

Random, undirected mutation is a major driving force for change in nature. In the protein universe, selection is realized through function: mutations leading to loss of function are rarely observed. As protein structure determines protein function, it is also subjected to evolutionary selection. Most problematic single nucleotide polymorphisms (SNP) that alter the amino acid sequence (non-synonymous SNPs) appear to impact the stability of protein structure (Yue *et al*., 2005; Yue *et al*., 2006).

Helices and strands constitute the major macromolecular building blocks of all 'well-ordered' proteins (Benner *et al*., 1997; Kabsch and Sander, 1983; Levitt and Chothia, 1976; Morea *et al*., 1998; Pauling and Corey, 1951a; Pauling and Corey, 1951b). The particular 3D structure of a protein is assumed to correspond to the global minimum free energy and hence defines the unique fold of an amino acid polymer (Anfinsen and Scheraga, 1975; Dill, 1993; Karplus and Petsko, 1990; Levitt and Warshel, 1975; Liwo *et al*., 1999; Reva *et al*., 1995; Sippl, 1993). Another essential feature of protein structure is the unique interplay between well-ordered and flexible regions (Alexov and Gunner, 1997; Cavasotto and Abagyan, 2004; Claussen *et al*., 2001; Daniel *et al*., 2003; Gu *et al*., 2006; Morea *et al*., 2000; Radivojac *et al*., 2004; Schlessinger *et al*., 2006). One particular aspect of this interplay is that between what we may loosely refer to as 'order' and 'disorder' (Dunker and Obradovic, 2001; Dunker *et al*., 2008; Radivojac *et al*., 2004; Uversky, 2003).

Many proteins have regions that remain 'unstructured' unless bound to a substrate: they do not adopt a unique stable conformation in isolation. Such regions are also referred to as *intrinsically disordered* or simply as *disordered*. Our operational definition for this vague term is: *we consider as disorder whatever is predicted as such*. Proteins with long-disorder regions have unique biophysical traits that enable the binding to different substrates, often at different cellular conditions (Wright and Dyson, 2009). Very long regions without regular secondary structure (loosely referred to as 'loops') may resemble disorder (Liu *et al*., 2002); nevertheless, we can clearly distinguish between disorder-like and well-structured loops (Schlessinger *et al*., 2007a; Schlessinger *et al*., 2009). Disorder is an important 'building block' for the increase in complexity in the evolution from unicellular prokaryotes to multi-cellular eukaryotes.

Our two hypotheses were: (i) we assumed that regular secondary structure is difficult to maintain evolutionarily, i.e. single residue mutations are likely to impact helices and strands and that we would lose regular secondary structure and transit into 'loopy' polypeptide chains with increasing random mutations away from the native state. (ii) We assumed, furthermore, that disordered regions provide a means to become robust against mutations because most mutations would rather increase than decrease disorder by increasing the non-regular secondary structure. Here, we present results that falsify both hypotheses as clearly as possible without investing tens of millions of dollars.

2 METHODS

2.1 Datasets

We used protein sequences from two databases for the *in silico* mutation. First, we assessed the robustness of secondary structure through globular proteins from the Protein Data Bank (PDB) (Berman *et al*., 2000). Secondly, we assessed the robustness of disordered regions through proteins from DisProt (Vucetic *et al*., 2005) (version 4.9). We applied UniqueProt (Mika and Rost, 2003) to reduce the redundancy in both sets filtering at a sequence similarity threshold of HVAL > 10 (Rost, 1999; Sander and Schneider, 1991) (this corresponds to ∼30% pairwise sequence identity—PIDE—for alignments over 250 residues). The redundancy-reduced sets comprised 1369 (PDB) and 374 (DisProt) proteins.

For each of the two datasets (PDB and DisProt), we also created random sequences that had the same amino acid composition, same length distribution and same number of sequences as the natives. The random sets served as convergence control: if we mutate enough to 'lose all memory' (convergence), the random sets will not differ from the mutated sets.

To shed light on potential biases from the chosen databases, we additionally predicted the secondary structure in 33 812 proteins, representing the entire human proteome as taken from RefSeq 2006.

Finally, we sub-sampled a set of sequences from the PDB set with the same size, amino acid and length distribution as that of the DisProt set to examine the ability of ordered proteins to retain or lose their ordered state.

2.2 Mutation protocol

We gradually mutated native protein sequences into quasi-random strings of amino acids by the following iterative procedure.

2.2.1 One mutation step It consisted of two moves: (i) select a particular residue position, i.e. site in the sequence to mutate, and (ii) mutate the amino acid X at that position with amino acid Y with the probability p_{XY} (X=Y). For technical reasons (lack of CPU because after each step we have to apply several prediction methods), we repeat these two moves $N/10$ times (N number of residues in the protein). Effectively, we thereby touch 10% of all residues in one mutation step.

2.2.2 Sixty-nine mutation steps We carried out 69 mutation steps (with $69 \times N/10$ mutations) for each protein. Any other, sufficiently large, number would have worked. We chose 69 because we had reached convergence in all the cases that we looked at in detail after 65 steps.

Effectively, we applied a Markovian-like model for evolution, i.e. assuming that each residue mutates independently of all others and that the mutation depends only on the amino acid type. We applied three alternative substitution schemes: (i) we mutated according to the PAM120 probability (Dayhoff, 1978). (ii) PAM120 is valid for great evolutionary distance. In order to also cover closer relations, we also implemented BLOSUM62 (Henikoff and Henikoff, 1992). (iii) Finally, we took the underlying amino acid distribution in the database (PDB, DisProt—ordered/disordered regions in DisProt not distinguished) as substitution probabilities. Note that for the most PAM120 and BLOSUM62 mutations, the most likely 'mutation step' was the maintenance of the current amino acid as the diagonals are typically highest in these matrices. We did not consider mutations that led to insertions or deletions. BLOSUM62 and PAM120 behaved identically with respect to our results. For readability, we confined the BLOSUM62 results to the Supplementary Material.

2.2.3 Single trajectory versus ensemble The 'mutation path' for each native sequence constitutes a single unique trajectory in the space of all possible mutations. We created five different such single paths (five different mutants) in order to investigate the divergence from the native of an ensemble of evolutionary paths. From these five, we compiled a consensus by per-residue averaging over each of the five predictions (secondary structure/disorder). Note that by default, we reported the results for single trajectories and added the ensemble comparison only where explicitly stated.

2.3 Secondary structure

We predicted secondary structure through PROFsec (Rost, 2005). Secondary structure prediction methods improve when using evolutionary information (Liu and Rost, 2001; Rost, 1996; Rost and Sander, 1993). Without this information, PROFsec reaches a sustained single-sequence level of ∼68% three-state per-residue accuracy (Q_3 is the percentage of residues predicted correctly in one of the three states helix, strand and other). We had to use this single-sequence mode to monitor the effect of point mutations. Prediction mistakes might invalidate the generality of our findings. One way in which we addressed this concern was by monitoring the parameters that we plotted for our mutants also for the experimental observations from the native proteins as taken from DSSP (Kabsch and Sander, 1983) with the usual conversion of eight into three 'states' (Andersen *et al*., 2002; Rost, 1996; Rost and Sander, 1993). For each mutation step (i.e. after each step of 10% change), we monitored the sequence similarity compared with the native sequence, the relative content of residues predicted in helix and strand and the average length of predicted helices and strands.

2.4 Disordered regions

We predicted disordered regions by three methods: IUPred (Dosztányi *et al*., 2005), MD (Schlessinger *et al*., 2009) and VSL2 (Obradovic *et al*., 2005; Peng *et al*., 2006) and compared the predictions to the experimental annotations in DisProt. IUPred has three options (*long*, *short* and *glob*); we chose *short* for short and *long* for long disorder. MD (Meta Disorder predictor) combines independent methods through machine learning. We used it without alignments. VSL2 is a collection of eight methods. We used the VSL2B variant that uses only single sequences as input.

The three methods focus on different aspects of disorder and have different strengths and weaknesses. We did not combine methods and, for simplicity, focused only on IUPred. The results from the other methods that were crucial to rule out method-specific findings are given in the Supplementary Material. We chose IUPred because it is accurate, fast and set up to work only with single sequences.

For each mutation step (i.e. after each step of 10% change), we monitored sequence similarity to native, the relative content of residues predicted in short/long-disordered regions and the length of the regions (SOM).

2.5 Box plots to present results

Box plots (McGill *et al*., 1978; Tukey, 1977) present our results concisely. The lower and upper box edges depict the first and third quartile, respectively. The length of a box is the interquartile range of the distribution. The bold bar inside the box represents the median, while dashed lines reach to the most extreme data point that is no more than 1.5 times the interquartile range away from the upper or lower box edge. Average (mean) values are connected through solid lines and intersect with box plots.

Median and mean are related to the protein level, i.e. summarize the specific feature of all sequences that fall within the same interval of PIDE.

3 RESULTS AND DISCUSSION

3.1 Secondary structure surprisingly robust

Comparisons of pairs of evolutionarily related protein structures reveal two major results (Abagyan and Batalov, 1997; Chothia and Lesk, 1986; Chung and Subbiah, 1996; Sander and Schneider, 1991): first, the less similar their sequences, the less similar their 3D structures [as well as their secondary structures (Rost *et al.*, 1994; Rost *et al.*, 1997)]; and second, the transition from the regime of 'similar structure' to 'non-similar structure' is highly non-linear and characterized by sigmoids indicative of phase transitions in physics. Our mutation protocol yielded a very different outcome.

Secondary structure diverged to almost random levels over the course of our mutation protocol. We compared this divergence to what is observed between naturally occurring homologues. Towards this end, we used the HSSP database (Sander and Schneider, 1991) and compared homologues at the corresponding levels of PIDE (Supplementary Fig. SOM_5). The change of secondary structure on random mutation was much more dramatic than that for homologous proteins (Fig. 1A), e.g. at 30%, PIDE natural homologues still had levels of Q_3 ~63%, while the random mutants reached Q_3 ~45% (Supplementary Fig. SOM_5). This result is not surprising: evolution *feels* the pressure to enrich neutral mutations, i.e. those that do not alter structure, while no such incentive was built into our *in silico* mutation protocol. Nevertheless, secondary structure was surprisingly robust under mutation. The consensus over ensembles of five different mutation trajectories (Fig. 1C and D) diverged much more dramatically from wild type than any single mutant (Fig. 1A and B).

Another important difference between our *in silico* mutation and natural evolution pertained to the shape of the transition: instead of a sigmoidal phase transition, we observed an almost linear transition from native wild-type to almost random mutant. This was true for both the single trajectory (Fig. 1A) and the ensemble (Fig. 1C), although the signal was clearer for the ensemble.

We observed that some regions did not alter secondary structure even at the end of our protocol at which the mutant was as similar to the wild type as to any other sequence in our dataset (Fig. 1B). For the ensemble, in contrast, the consensus secondary structure had changed almost completely from the native (Fig. 1D). Nevertheless, the Q3 levels converged to the same level in both cases.

3.2 Helix and strand intrinsic to random sequences

Our most surprising finding was that neither the overall content (Fig. 2A and B) nor the length (Fig. 2C and D) of *predicted* helices and strands was altered during the course of our mutation protocol. The average helix content remained ~30%, whereas the average strand content around 20%; the average helix was about 10 residues long (2–3 helix turns), and the average strand extended over about five residues. In other words, regular secondary structure was predicted to be robust under extreme mutation. In this respect, we observed no significant difference between choosing mutations according to the background distribution and PAM120, although the latter tends to follow the evolutionarily more accepted mutations (mutations according to BLOSUM62 gave similar results Supplementary Fig. SOM_6).

After the 69 mutation steps (Section 2), we reached a point at which the mutant was as similar to the native as to any other sequence. This was reflected by the similarity in the prediction of helix/strand content/length between the final mutant and randomly created sequences (Fig. 2: two rightmost bars almost identical).

Our results were based on predictions rather than on observations. Prediction methods make mistakes. One might hypothesize that rather than shedding light on protein features, our results are caused by those prediction mistakes. As no large-scale experiments establish structure for random sequences, we cannot refute this view. However, we could provide evidence that prediction mistakes might

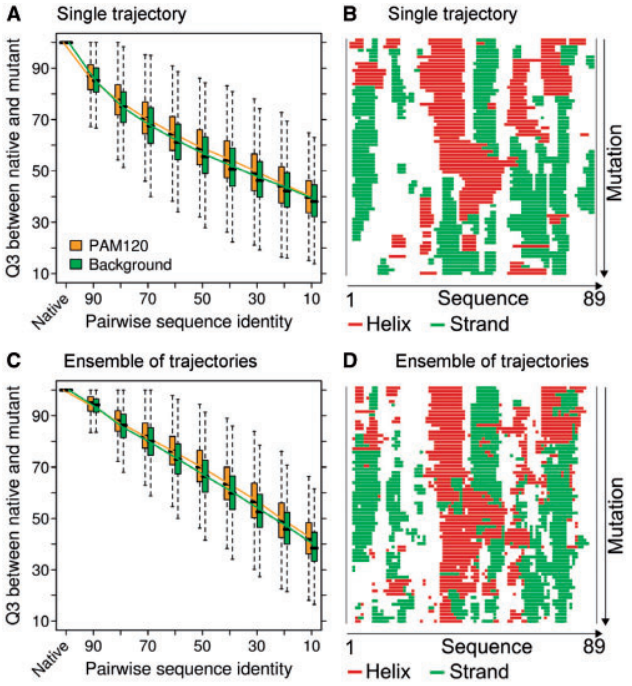

Fig. 1. Secondary structure changes proportional to sequence. (**A** and **C**) For decreasing pairwise percentage sequence identity (*x*-axis, PIDE), we monitored the similarity between secondary structure predictions (Q_3, i.e. percentage of residues identical in one of the three states helix, strand and other) for native and for mutant (yellow: mutations according to PAM120, green: according to background distribution, Section 2). (A and B) show results for a single trajectory, (C and D) the consensus over an ensemble of five trajectories (Section 2). Box plots reflect the range of the distribution (Section 2); median values are marked by horizontal bars and mean values are connected by dotted lines. For instance, at ~90% pairwise sequence identity, ~88% of the residues are predicted in the same secondary structure as the native; for the ensemble, this value is slightly higher (leftmost bars in A and C). The curves converge nearly linearly towards values ~35% corresponding to random. **(B** and **D)** For one particular example (PDB identifier 1a2s chain A), we display the actual secondary structure predictions for each mutant: native on top; each row marks one of the 69 mutation steps (Section 2); mutation by PAM120. The top (B) is for one single mutation trajectory, the bottom (D) for an ensemble of five trajectories. One observation stands out and is representative for all such plots that we looked at: blocks of regular secondary appear to be more robust under mutation than the actual type of secondary structure, i.e. helices flip to strands and vice versa and this happens more often than the transitions helix→other and strand→other. Borders are much more 'fluid' for the ensemble (D) than for a single mutation trajectory (B).

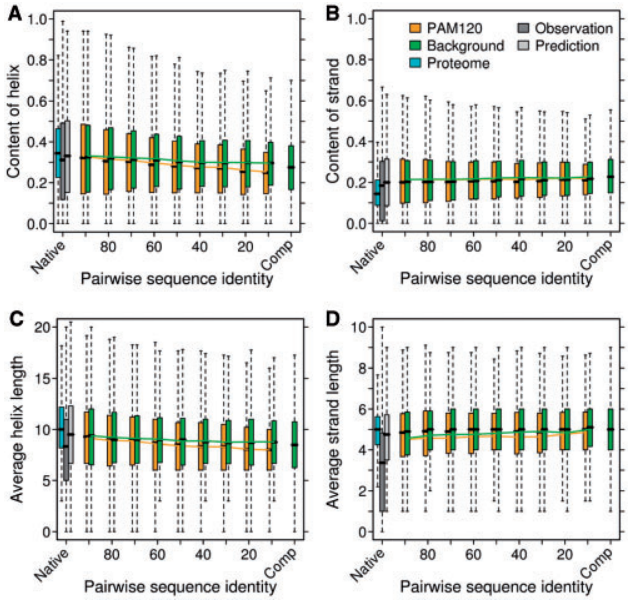

Fig. 2. Content and length of regular secondary structure unchanged. Box plots and coloring as in Figure 1. Change of regular secondary structure on mutation given by the composition of predicted helix (**A**) and strand (**B**), as well as the average lengths of predicted helices (**C**) and strands (**D**). The second and third bar on the left in (A) and (B) compare predictions (light gray) with observations (taken from DSSP, dark gray) for the PDB dataset; the first bar on the left in (A) and (B) indicates the degree to which the predictions differ for the PDB dataset (dark gray) and for a set of all human proteins (light blue). The right-most green bars mark the predictions for randomly assembled sequences (Section 2, labeled as 'Comp'). Overall, neither the length nor the content of regular secondary structure appears to differ between native and random.

not matter for the aspects of structure that we monitored. In fact, by the measures that we used to report our results, predictions and observations were almost identical (Fig. 2: left gray bars in each panel). The precise levels of helix/strand content and length differed indeed more between different datasets (PDB subset versus entire set of human proteins) than between observation and prediction for any set for which we have experimental information. In other words, prediction mistakes appeared not to matter for all the proteins for which we could verify this statement.

Our findings that random and wild-type sequences were predicted to have similar content of regular secondary structure along with the observation that mistakes in predicting this were negligible suggest that the formation of helices and strands is an intrinsic feature of amino acid sequences. Neither helices nor strands were predicted to be significantly shortened during our drastic *in silico* mutation protocol. Note that this is not a consequence of the fact that PROFsec is trained to predict a particular length distribution, because predicted length distributions deviate substantially between all-helical and coiled-coil proteins. The maintenance of such regular secondary structure elements would then appear to come at seemingly low costs, i.e. mutations that are neutral with respect to structure might be more likely than might have been anticipated. Finally, we verified that the reliability of the predictions did not change during mutation (Supplementary Fig. SOM_10).

3.3 Long regions of disorder sensitive, short not

Arguably, there are two different regimes of disorder (Dosztányi *et al*., 2005; Liu *et al*., 2002; Obradovic *et al*., 2005; Peng *et al*., 2006; Schlessinger *et al*., 2007b; Schlessinger *et al*., 2009): very short and very long regions. No threshold distinguishes between these two regimes in a biophysically meaningful way.

In particular, there likely exists an intermediate range that might belong to both regimes. Here, we followed the typical 'convention' in the field and defined as short disorder regions with eight or less consecutive residues and as long disorder regions with 30 or more consecutive residues. Thereby, we ignored the uncertain regime in between these two extremes. In order to establish that our results did not crucially depend on the particular threshold, we also tested other thresholds for long disorder, namely 20, 40 and 50. We found that the trend of loss during *in silico* mutation is independent of the chosen cut-off and is even clearer for larger thresholds (40 and 50) (Supplementary Fig. SOM_09).

First, we observed that regions of short disorder behaved like regular secondary structure in that their content (Fig. 3B, D and F; Supplementary Fig. SOM_2D and E) and length (Supplementary Fig. SOM_2A–C) did not alter on mutation. In stark contrast was the result for long regions with predicted disorder gradually diminished over the course of our mutation protocol (Fig. 3A, C and E; by definition a prediction of 29 disordered residues for some mutant implies that for that mutant the long disordered region seemingly 'disappeared', e.g. Fig. 3E middle; Supplementary Fig. SOM_1). The loss on mutation was much more dramatic for mutations according to PAM120 (yellow in Fig. 3C) than for those according to the background distribution (green in Fig. 3C). This is understandable because disordered regions are abundant in polar residues, and these are more likely to be chosen if mutation probability is 'skewed' toward this abundance. Put differently, PAM120-driven mutations drifted toward sequences that resembled regular well-structured proteins and as such had no disorder, while background-driven mutations yielded sequences that were as abundant in disorder as the native wild types and therefore had many long regions with predicted disorder.

The actual numbers in terms of content of predicted long disorder decreased from ∼18% for the native to ∼9% for the final mutant by using the background mutation protocol (Fig. 3C, green). This reflected the fact that a considerable fraction of the residues in our DisProt dataset was polar: for mutations according to PAM120 (Fig. 3C, yellow) or BLOSUM62 (Supplementary Fig. SOM_7), the content dropped to 0. However, at this level of mutations, almost no single residue predicted as long disorder in the native was predicted as disorder in the mutant (Fig. 3A). For some, this might appear to PAM120.

Studies of particular mutation paths revealed that long disorder might just appear to vanish *suddenly* (Fig. 3E). This was partially a threshold issue: assume a region with 35 consecutive 'disordered' residues and assume the mutant loses three on each side (six in total); we will no longer consider this as long disorder (35–6 <30). This also explains how additional mutations may *recover* the long disorder (Fig. 3E: after solid block of red bars, suddenly one mutant has disorder again as seen by a single bar below this block).

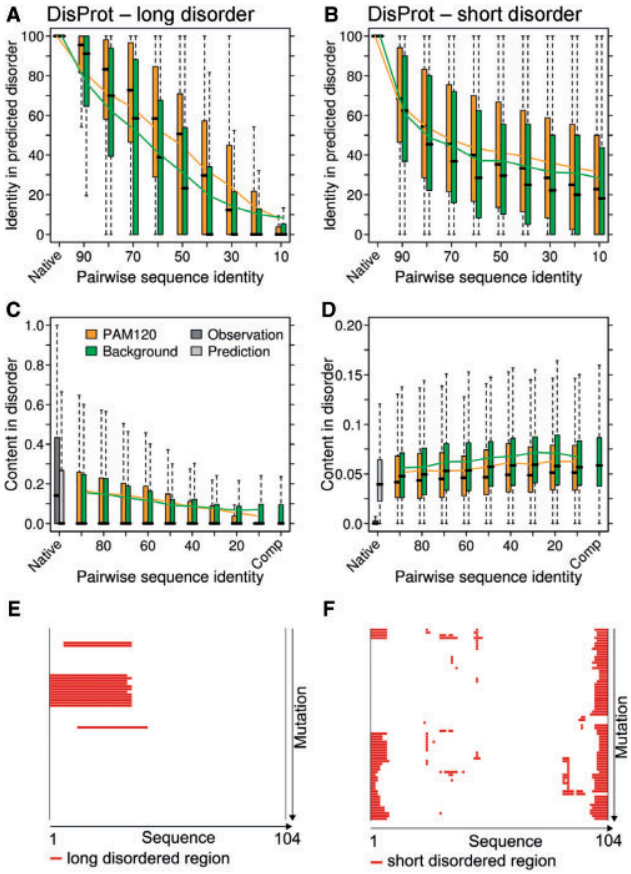

Fig. 3. Predicted long disorder changes rapidly. Panels on the left show results for long regions of disorder (30 or more consecutive residues), those on the right for short regions (less than eight). The top panels (**A** and **B**) demonstrate how much the predictions of disorder changed over the course of mutations (*y*-axis: residues predicted identical as disorder between native and mutant as percentage of disorder predicted in native). Disorder predictions differ much more rapidly from native than do secondary structure predictions, and much more for long (A) than for short (B) disorder. The relative content of residues in predicted long (**C**) and short (**D**) disordered regions diverge differentially. The first two box plots for (C) depict the observed (dark gray) and predicted (light gray) disordered content in native sequences. Right box plots in both (C) and (D) show the disordered situation in the artificially created dataset sequences (Section 2, labeled as 'Comp'). For a representative example (DisProt identifier: DP 00006), the IUPred predictions for long (**E**) and short (**F**) disorder are shown for each mutant: native on top; each row marks 1 of the 69 PAM120 mutation steps (Section 2). Red lines mark predictions that fall into the threshold category ((30 or more/less than eight). Long disordered regions disappear (E) while especially short disorder remains at both termini, while re- and disappearing in the middle region during mutation (**F**).

Another observation reflects one of the important aspects when studying short disorder: a considerable fraction of the short disorder is predicted (and observed) near the protein termini (Fig. 3F). Short disorder 'comes and goes' during mutation (middle region in Fig. 3F). Although this effect is biologically relevant and dominates the study of disorder in otherwise well-ordered proteins (Bordoli *et al.*, 2007; Jin and Dunbrack, 2005), it again underlines the problem of not differentiating between long and short disorder.

Our analyses of regular secondary structure and disorder are based on very different datasets. PDB is biased in many ways (Liu and Rost, 2001), one of those pertains to disorder (Liu and Deber, 1999; Peng *et al.*, 2004). One reason simply is that proteins with disordered regions pose extreme challenges to structure determination (Burley *et al.*, 2008; Dunker *et al.*, 2008; Graslund *et al.*, 2008; Liu *et al.*, 2004; Nair *et al.*, 2009; Romier *et al.*, 2006). To address this difference, we predicted disorder also for the dataset of well-ordered proteins from the PDB. As expected, the level of both long and short disorder for both of those was very low (Supplementary Figs SOM_3 and 4); given the lack of disorder in these proteins, we could therefore not observe any significant difference between close-to-zero in the wild type and close-to-zero in the mutants.

IUPred is arguably one of the best disorder prediction methods (Bordoli *et al.*, 2007; Le Gall *et al.*, 2007; Schlessinger *et al.*, 2007b; Schlessinger *et al.*, 2009; Shimizu *et al.*, 2007); however, it is still only one of many and it has specific strengths and weaknesses. Therefore, we also predicted disorder with two other state-of-the-art prediction methods, namely VSL2 (Obradovic *et al.*, 2005; Peng *et al.*, 2006) and MD (Schlessinger *et al.*, 2009). Although the predictions for those two differed slightly from those for IUPred, by the measures we reported here, they revealed exactly the same trend: while predicted long disorder disappeared on mutation, the content and length distribution of predicted short disorder remained largely unaffected by the mutation.

We addressed the impact of incorrect predictions by randomly introducing errors. At any significant error rate, long disorder disappeared in the native. This highlights the high prediction accuracy of today's methods. For short disorder, the added error did not alter the content over the course of our mutation protocol (Supplementary Fig. SOM_8).

As short and long disorders have different physical traits, we need length thresholds. However, we can drop these thresholds while monitoring the disappearance of disorder. Toward this end, we began with all native regions longer than N (chosen in steps of between 20 and 50), and monitored the percentage of disorder predicted after mutation irrespective of the length of the predicted regions. We found that long disordered regions indeed get decomposed into shorter ones and that disorder disappears throughout (Supplementary Figs SOM_11 and 12).

4 CONCLUSIONS

We addressed the general question whether or not well-ordered regular secondary structure and disordered regions sustain random mutations. Is it likely or unlikely that any mutation affects this particular coarse-grained feature of protein structure (and through it's function)? Do random sequences have different content in secondary structure and disorder than native proteins that have evolved to satisfy many constraints? Our analysis clearly suggests two different answers for regular secondary structure and long disorder. On the one hand, the maintenance of regular secondary structure might not be too challenging because its formation appears to be an intrinsic feature of random sequences. It, therefore, appears surprisingly likely to transit from helix to strand and back. In fact, this is exactly what we dynamically observed during the course of our mutations (Fig. 4). On the other hand, regions of long disorder do not appear to be robust under mutation. Random changes likely disrupt this feature that thereby appears volatile and unique.

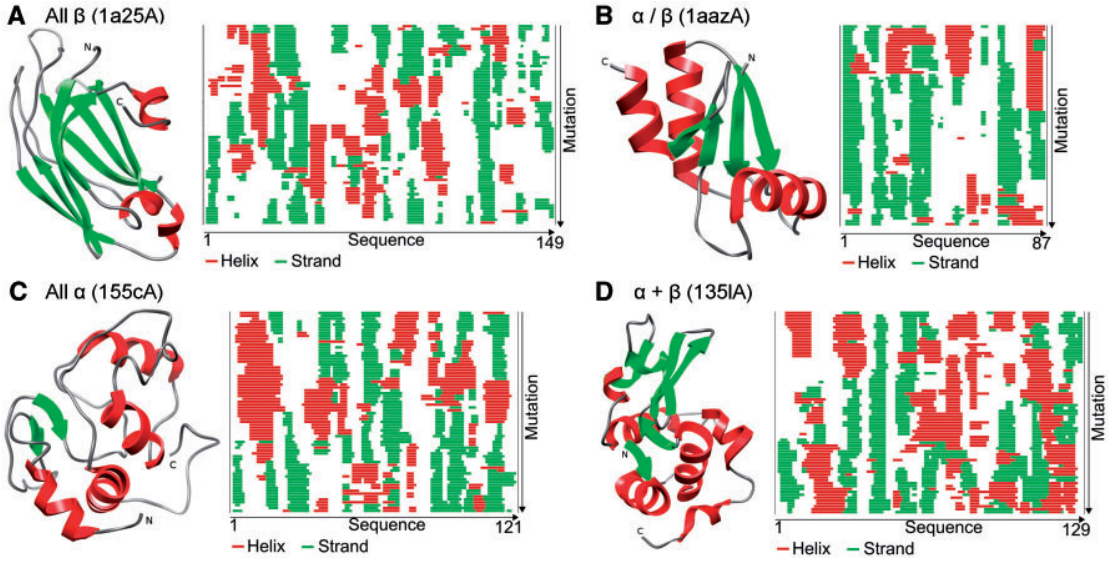

Fig. 4. Examples of proteins with mutation trajectories. For each of the four main SCOP classes (Murzin *et al*., 1995), we randomly picked one representative short enough to fit into the space here. Ribbon plots were generated by Chimera (Pettersen *et al*., 2004) [red: helix, green: strand, according to DSSP (Kabsch and Sander, 1983)]. (**A**–**D**) In each of the four panels, the ribbon diagram for the native is on the left, and on the right are the 69 mutation trajectories (top: native, degree of mutation decreases downwards; mutations according to PAM120, Section 2). The sequence runs from the most N-terminal residues (labeled '1') to the most C-terminal ones. Note that although we show only single trajectories, rather than ensemble averages here, almost no helix or strand withstands the mutation protocol to the end.

This has important impact on how we picture the role of long disorder in proteins: it is not 'easy' to acquire. Prokaryotes have only ∼10–25% of the disorder observed in multi-cellular eukaryotes (Dunker *et al*., 2008; Ekman *et al*., 2005; Liu *et al*., 2002; Oldfield *et al*., 2005; Romero *et al*., 2004; Schlessinger *et al*., 2009; Ward *et al*., 2004). Our observation of how volatile long disorder is provides another evidence for the importance of this feature for the transition from prokaryotes to eukaryotes.

Many SNPs that alter the protein sequence (nsSNPs) appear to be deleterious. Is this a bias in the experimental technique (more likely to be observed/reported if deleterious), or is it a genuine feature of proteins imposed by the sensitivity of protein structure to mutations? Although our work neither addresses nor answers this question, the surprising robustness of regular secondary structure might support the view that protein structure is more flexible and adaptable than the intricate details of the concert of interacting residues in protein 3D structures might suggest.

ACKNOWLEDGEMENTS

The authors would like to thank the following for valuable discussions: Zsuzsanna Dosztanyi (Eötvös Loránd University Budapest and Columbia University in the City of New York), Dietlind Gerloff (UCSC Santa Cruz), Marco Punta (Columbia University in the City of New York and TUM Munich), Reinhard Schneider (EMBL Heidelberg), Anna Tramontano (La Sapienza Rome and KAUST); the anonymous reviewers for very constructive and helpful suggestions that helped shaping this work; and also to all those who deposit their experimental data in public databases and to those who maintain these databases, in particular to those who contribute to PDB and DisProt.

Funding: National Institute of General Medical Sciences (NIGMS) at the National Institutes of Health (NIH) (grant number R01-GM079767).

Conflict of Interest: none declared.

REFERENCES

Abagyan,R.A. and Batalov,S. (1997) Do aligned sequences share the same fold? *J. Mol. Biol.*, **273**, 355–368.

Alexov,E.G. and Gunner,M.R. (1997) Incorporating protein conformational flexibility into the calculation of pH-dependent protein properties. *Biophys. J.*, **72**, 2075–2093.

Andersen,C.A.F. *et al*. (2002) Continuum secondary structure captures protein flexibility. *Structure*, **10**, 175–184.

Anfinsen,C.B. and Scheraga,H.A. (1975) Experimental and theoretical aspects of protein folding. *Adv. Prot. Chem.*, **29**, 205–300.

Benner,S.A. *et al*. (1997) Bona fide predictions of protein secondary structure using transparent analyses of multiple sequence alignments. *Chem. Rev.*, **97**, 2725–2844.

Berman,H. *et al*. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.

Bordoli,L. *et al*. (2007) Assessment of disorder predictions in CASP7. *Prot. Struct. Funct. Genet.*, **69**(Suppl. 8), 129–136.

Burley,S.K. *et al*. (2008) Contributions to the NIH-NIGMS protein structure initiative from the PSI production centers. *Structure*, **16**, 5–11.

Cavasotto,C.N. and Abagyan,R.A. (2004) Protein flexibility in ligand docking and virtual screening to protein kinases. *J. Mol. Biol.*, **337**, 209–225.

Chothia,C. and Lesk,A.M. (1986) The use of sequence homologies to predict protein structures. In Robert,F. and Mark,Z. (eds) *Computer Graphics and Molecular Modeling*. Cold Spring Harbor Laboratory, New York, pp. 33–37.

Chung,S.Y. and Subbiah,S. (1996) A structural explanation for the twilight zone of protein sequence homology. *Structure*, **4**, 1123–1127.

Claussen,H. *et al*. (2001) FlexE: efficient molecular docking considering protein structure variations. *J. Mol. Biol.*, **308**, 377–395.

Daniel,R.M. *et al*. (2003) The role of dynamics in enzyme activity. *Annu. Rev. Biophys. Biomol. Struct.*, **32**, 69–92.

Dayhoff,M.O. (1978) *Atlas of Protein Sequence and Structure*. National Biomedical Research Foundation, Silver Spring, MD, pp. 345–358.

Dill,K.A. (1993) Folding proteins: finding a needle in a haystack. *Curr. Opin. Struct. Biol.*, **3**, 99–103.

Dosztányi,Z. *et al*. (2005) The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. *J. Mol. Biol.*, **347**, 827–839.

Dunker,A.K. and Obradovic,Z. (2001) The protein trinity-linking function and disorder. *Nat. Biotechnol.*, **19**, 805–806.

Dunker,A.K. *et al*. (2008) Function and structure of inherently disordered proteins. *Curr. Opin. Struct. Biol.*, **18**, 756–764.

Ekman,D. *et al*. (2005) Multi-domain proteins in the three kingdoms of life: orphan domains and other unassigned regions. *J. Mol. Biol.*, **348**, 231–243.

Graslund,S. *et al*. (2008) Protein production and purification. *Nat. Methods*, **5**, 135–146.

Gu,J. *et al*. (2006) Wiggle-predicting functionally flexible regions from primary sequence. *PLoS Comput. Biol.*, **2**, e90.

Henikoff,S. and Henikoff,J.G. (1992) Amino acid substitution matrices from protein blocks. *Proc. Natl Acad. Sci. USA*, **89**, 10915–10919.

Jin,Y. and Dunbrack,R.L. Jr (2005) Assessment of disorder predictions in CASP6. *Proteins*, **61**(Suppl. 7), 167–175.

Kabsch,W. and Sander,C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.

Karplus,M. and Petsko,G.A. (1990) Molecular dynamics simulations in biology. *Nature*, **347**, 631–639.

Le Gall,T. *et al*. (2007) Intrinsic disorder in the Protein Data Bank. *J. Biomol. Struct. Dyn.*, **24**, 325–342.

Levitt,M. and Chothia,C. (1976) Structural patterns in globular proteins. *Nature*, **261**, 552–558.

Levitt,M. and Warshel,A. (1975) Computer simulation of protein folding. *Nature*, **253**, 694–698.

Liu,J. *et al*. (2004) Automatic target selection for structural genomics on eukaryotes. *Prot. Struct., Funct., Bioinform.*, **56**, 188–200.

Liu,J. and Rost,B. (2001) Comparing function and structure between entire proteomes. *Protein Sci.*, **10**, 1970–1979.

Liu,J. *et al*. (2002) Loopy proteins appear conserved in evolution. *J. Mol. Biol.*, **322**, 53–64.

Liu,L.P. and Deber,C.M. (1999) Combining hydrophobicity and helicity: a novel approach to membrane protein structure prediction. *Bioorg. Med. Chem.*, **7**, 1–7.

Liwo,A. *et al*. (1999) Protein structure prediction by global optimization of a potential energy function. *Proc. Natl Acad. Sci. USA*, **96**, 5482–5485.

McGill,R. *et al*. (1978) Variations of box plots. *Am Statistician*, **32**, 12–16.

Mika,S. and Rost,B. (2003) UniqueProt: creating representative protein sequence sets. *Nucleic Acids Res.*, **31**, 3789–3791.

Morea,V. *et al*. (1998) Protein structure prediction and design. *Biotechnol. Annu. Rev.*, **4**, 177–214.

Morea,V. *et al*. (2000) Antibody modeling: implications for engineering and design. *Methods*, **20**, 267–279.

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

Nair,R. *et al*. (2009) Structural genomics is the largest contributor of novel structural leverage. *J. Struct. Funct. Genomics*, **10**, 181–191.

Obradovic,Z. *et al*. (2005) Exploiting heterogeneous sequence properties improves prediction of protein disorder. *Prot. Struct., Funct., Genet*. **61**(Suppl. 7), 176–182.

Oldfield,C.J. *et al*. (2005) Comparing and combining predictors of mostly disordered proteins. *Biochemistry*, **44**, 1989–2000.

Pauling,L. and Corey,R.B. (1951a) Configurations of polypeptide chains with favored orientations around single bonds: two new pleated sheets. *Proc. Natl Acad. Sci.*, **37**, 729–740.

Pauling,L. and Corey,R.B. (1951b) The pleated sheet, a new layer configuration of polypeptide chains. *Proc. Natl Acad. Sci. USA*, **37**, 251–256.

Peng,K. *et al*. (2004) Exploring bias in the Protein Data Bank using contrast classifiers. *Pac. Symp. Biocomput.*, **9**, 435–446.

Peng,K. *et al*. (2006) Length-dependent prediction of protein intrinsic disorder. *BMC Bioinformatics*, **7**, 208.

Pettersen,E.F. *et al*. (2004) UCSF Chimera–a visualization system for exploratory research and analysis. *J. Comput. Chem.*, **25**, 1605–1612.

Radivojac,P. *et al*. (2004) Protein flexibility and intrinsic disorder. *Protein Sci.*, **13**, 71–80.

Reva,B.A. *et al*. (1995) Constructing lattice models of protein chains with side groups. *J. Comput. Biol.*, **2**, 527–535.

Romero,P. *et al*. (2004) Natively disordered proteins : functions and predictions. *Appl. Bioinform.*, **3**, 105–113.

Romier,C. *et al*. (2006) Co-expression of protein complexes in prokaryotic and eukaryotic hosts: experimental procedures, database tracking and case studies. *Acta Crystallogr. D Biol. Crystallogr.*, **62**, 1232–1242.

Rost,B. (1996) PHD: predicting one-dimensional protein structure by profile based neural networks. *Methods Enzymol.*, **266**, 525–539.

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

Rost,B. (2005) How to use protein 1-D structure predicted by PROFphd. In Walker,J.M. (ed.), *The Proteomics Protocols Handbook*, Humana Press, pp. 875–901.

Rost,B. and Sander,C. (1993) Prediction of protein secondary structure at better than 70% accuracy. *J. Mol. Biol.*, **232**, 584–599.

Rost,B. *et al*. (1994) Redefining the goals of protein secondary structure prediction. *J. Mol. Biol.*, **235**, 13–26.

Rost,B. *et al*. (1997) Protein fold recognition by prediction-based threading. *J. Mol. Biol.*, **270**, 471–480.

Sander,C. and Schneider,R. (1991) Database of homology-derived protein structures and the structural meaning of sequence alignment. *Prot. Struct. Funct. Genet.*, **9**, 56–68.

Schlessinger,A. *et al*. (2007a) Natively unstructured loops differ from other loops. *PLoS Comput. Biol.*, **3**, e140.

Schlessinger,A. *et al*. (2007b) Natively unstructured regions in proteins identified from contact predictions. *Bioinformatics*, **23**, 2376–2384.

Schlessinger,A. *et al*. (2009) Improved disorder prediction by combination of orthogonal approaches. *PLOS ONE*, **4**, e4433.

Schlessinger,A. *et al*. (2006) PROFbval: predict flexible and rigid residues in proteins. *Bioinformatics*, **22**, 891–893.

Shimizu,K. *et al*. (2007) Predicting mostly disordered proteins by using structure-unknown protein data. *BMC Bioinformatics*, **8**, 78.

Sippl,M.J. (1993) Boltzmann's principle, knowledge based mean fields and protein folding. An approach to the computational determination of protein structures. *J. Comput.-Aided Mol. Des.*, **7**, 473–501.

Tukey,J.W. (1977) *Exploratory Data Analysis*. Addison-Wesley Pub. Co., Reading, MA.

Uversky,V.N. (2003) Protein folding revisited. A polypeptide chain at the folding-misfolding-nonfolding cross-roads: which way to go? *Cell Mol. Life Sci.*, **60**, 1852–1871.

Vucetic,S. *et al*. (2005) DisProt: a database of protein disorder. *Bioinformatics*, **21**, 137–140.

Ward,J.J. *et al*. (2004) Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *J. Mol. Biol.*, **337**, 635–645.

Wright,P.E. and Dyson,H.J. (2009) Linking folding and binding. *Curr. Opin. Struct. Biol.*, **19**, 31–38.

Yue,P. *et al*. (2005) Loss of protein structure stability as a major causative factor in monogenic disease. *J. Mol. Biol.*, **353**, 459–473.

Yue,P. *et al*. (2006) SNPs3D: candidate gene and SNP selection for association studies. *BMC Bioinformatics*, **7**, 166.

Supporting online material for: Protein secondary structure appears to be robust under *in silico* evolution while protein disorder appears not to be

Christian Schaefer, Avner Schlessinger & Burkhard Rost

Table of Contents for Supporting Online Material

Short description of Supporting Online Material (SOM)

The Supporting Online Material is not essential to support any of the major results and points of our manuscript. Instead, we provide additional data to support some of the minor. Most SOM data pertains to disorder predictions.

Fig. SOM_1 establishes that three different disorder prediction methods (IUPred, MD, VSL) give similar results and shows the overall content in disorder and the average length of disordered regions throughout the mutation protocol.

Fig. SOM_2 convenes the same message as Fig. SOM_1 but for short disorder.

Fig. SOM_3 and Fig. SOM_4: While Fig. SOM_1 and Fig. SOM_2 show the "*in silico* evolution" for proteins with disordered regions (taken from DisProt), Fig. SOM_3 and Fig. SOM_4 give the same data for proteins that are largely well-ordered in the sense that they yielded high-resolution experimental structures that have been deposited in the PDB.

Fig. SOM_5 depicts the differences in secondary structure predictions between naturally evolved homologues as taken from the HSSP database. These results are important in light of those presented in the main paper (Figs. 1-2) by showing how different the behavior is for divergence under evolutionary constraints (homologues: observed only what maintains function/structure) and divergence under random mutations.

Fig. SOM_6 and SOM_7 compare the impact of both PAM120 and BLOSUM62 mutations on variation in secondary structure and disorder.

Fig._SOM_8 shows the impact of randomly introduced errors in disorder predictions, i.e. the upkeep of short disorder during in-silico mutation even under high prediction error rates.

Fig. SOM_9 depicts that the general trend of long disorder, i.e. its loss, during in-silico mutation seems independent of the chosen cutoff.

Fig. SOM_10 shows that PROFsec's reliability stays quite constant during mutating away from the native state.

Fig. SOM_11 and SOM_12 show the behavior of 'raw' disorder without any length cutoff being applied.

Fig. SOM_1:

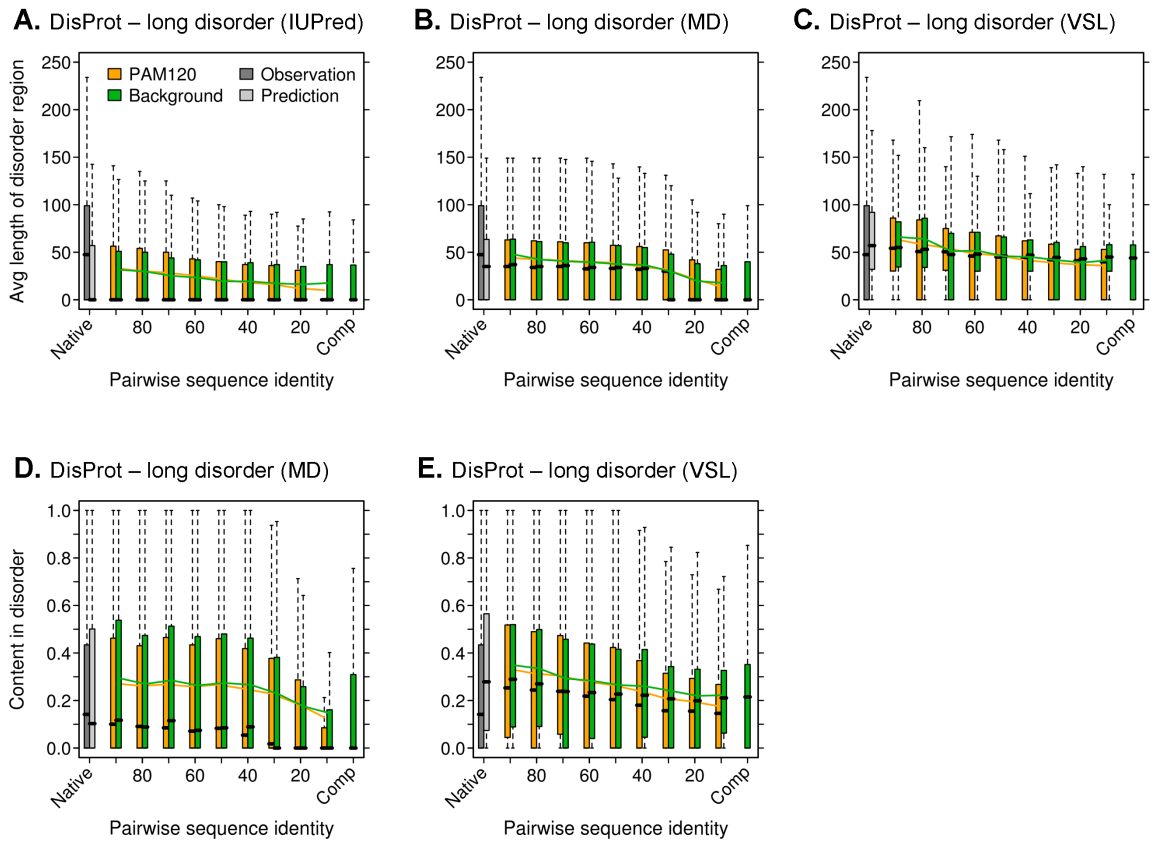

Fig. SOM_1: Mutation of long disorder in DisProt. (A-C) The average length of long disordered regions (≥30 residues) in DisProt proteins is decreasing over the course of our *in silico* mutation protocol (background green, PAM120 yellow, s. Methods) under all three disorder predictors. First two box plots (Methods) compare observation (dark gray) with prediction (light gray) in native protein sequences. The right-most green bars represent the randomly assembled sequences (Methods). **(D+E)** Content of long disorder drops in both mutation schemes (green, yellow lines), while the effect is more dramatic with MD (D) compared to VSL (E). Note that the situation with IUPred is shown in the original paper (Fig. 3A).

Fig. SOM_2:

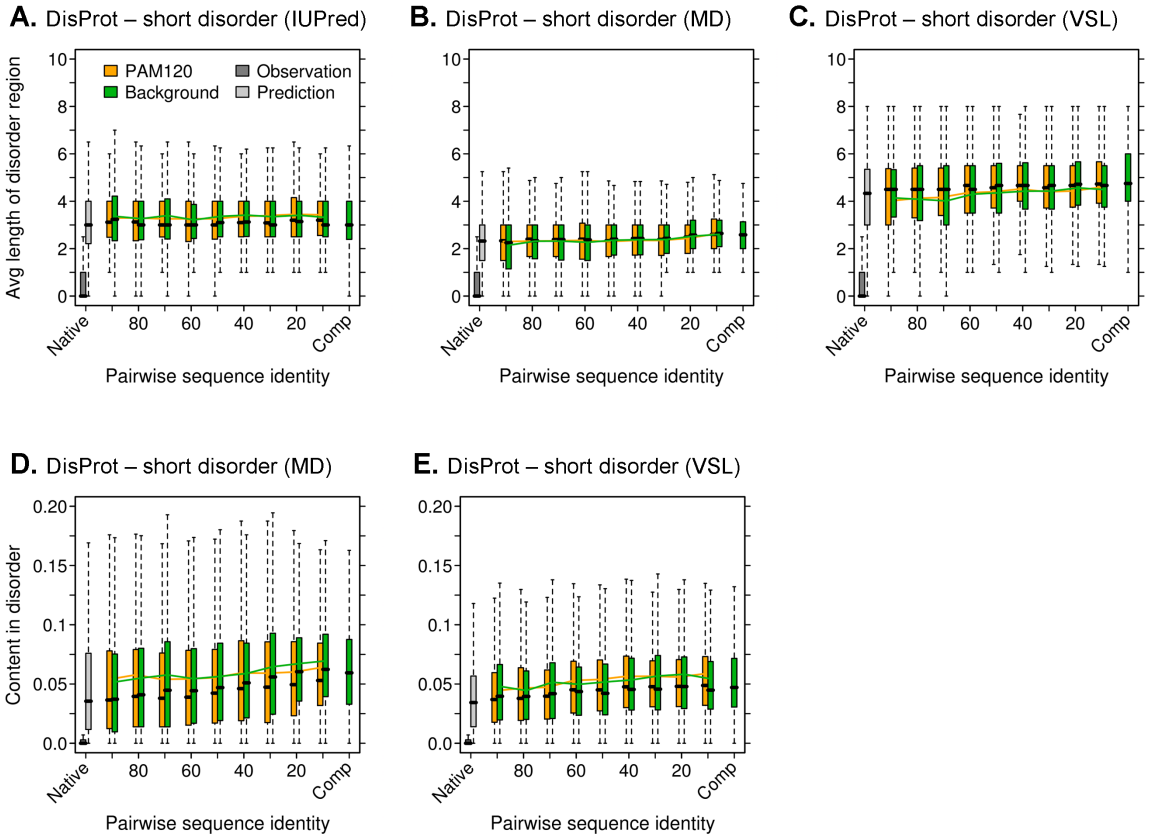

Fig. SOM_2: Mutation of short disorder in DisProt. Coloring and labels as in Fig. SOM_1. Difference: here we look only at short disorder (≤ 8 consecutive residues). **(A-C)** The average length of short disordered regions (≤8 residues) in DisProt proteins stays at nearly constant level for all three predictors. DisProt mainly contains long disorder hence the short bars around zero for observations (dark gray). **(D+E)** Content of short disorder also stays nearly constant for both MD (D) and VSL (E) predictions (IUPred in the original paper Fig. 3B).

Fig. SOM_3:

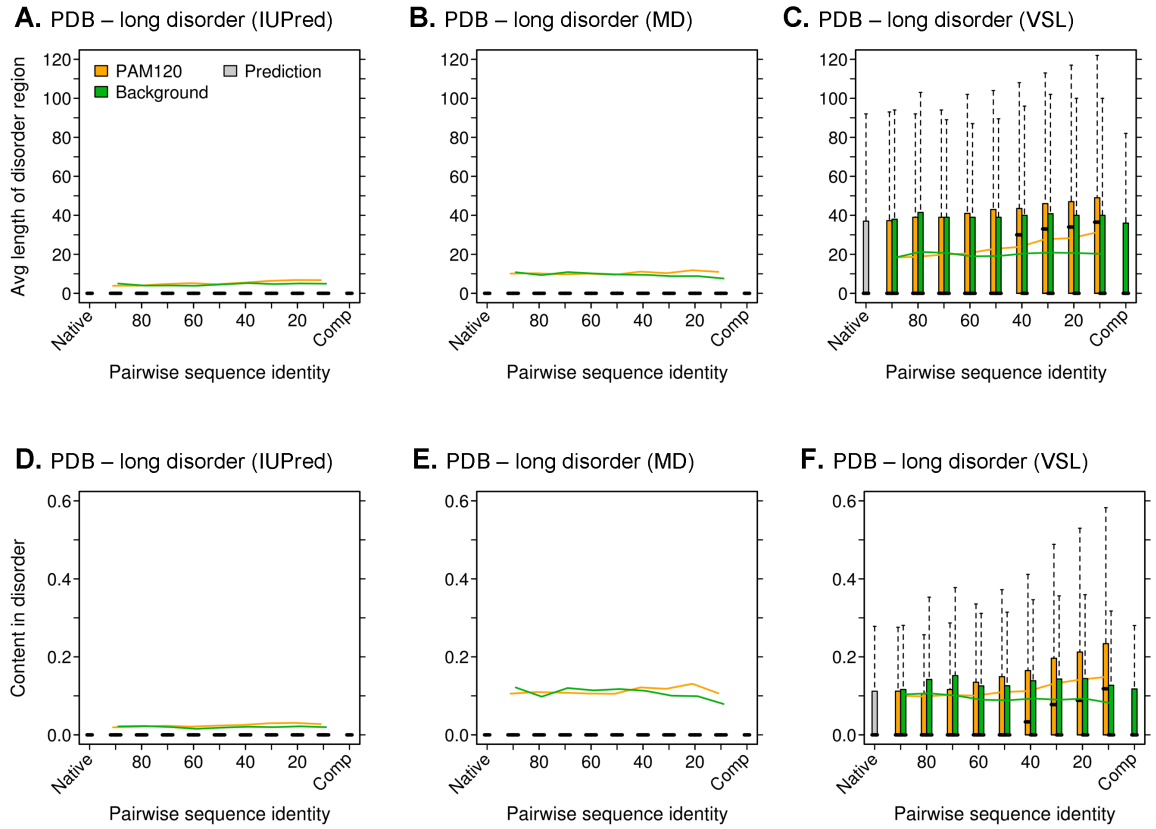

Fig. SOM_3: Mutation of long disorder in PDB. Coloring and labels as in Fig. SOM_1. Difference: here we look long disorder for proteins from the PDB. **(A-C)** Average length of long disorder in PDB stays at zero (degenerated box plots; higher mean values due to outliers) for both IUPred (A) and MD (B) while VSL (C) shows slightly elevated mean levels; PAM120 mutation increases the length. **(D-F)** Content of long disorder also stays at zero level for IUPred (D) and MD (E) while VSL (F) shows elevated levels.

Fig. SOM_4:

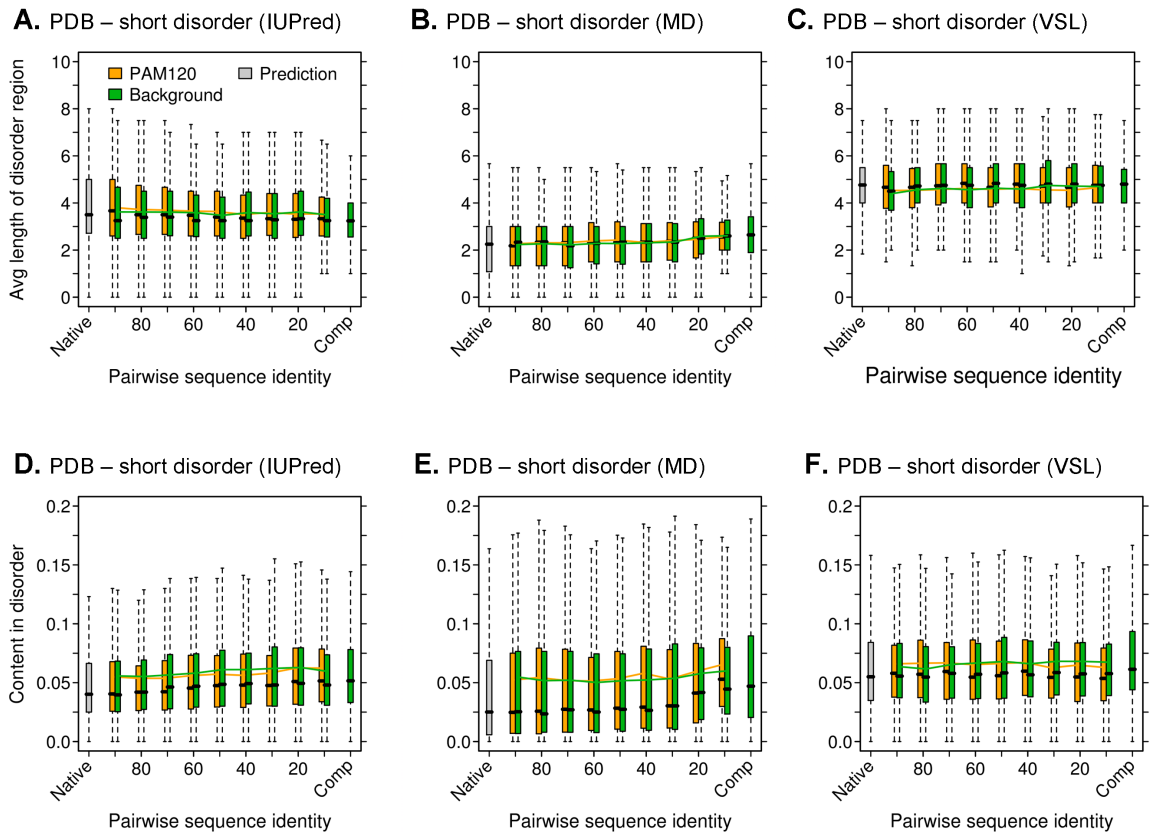

Fig. SOM_4: Mutation of short disorder in PDB. Coloring and labels as in Fig. SOM_1. Difference: here we look only at short disorder (≤ 8 consecutive residues) for proteins from the PDB. Average length **(A-C)** and content **(D-F)** of short disorder in PDB stays on constant level for all three predictors and are comparable to those in DisProt.

Fig. SOM_5:

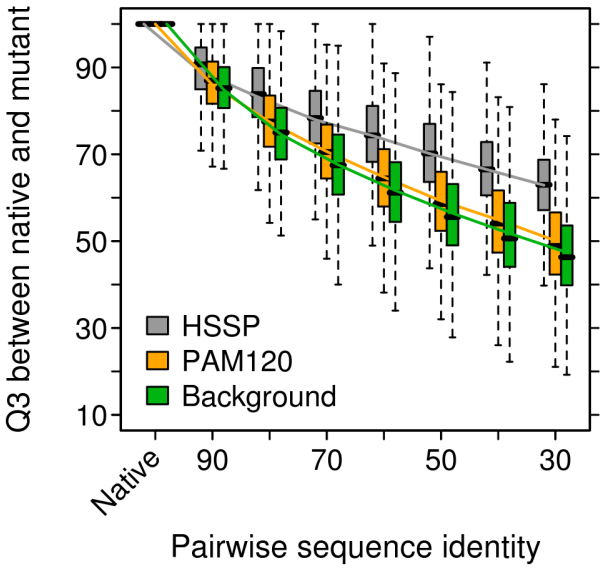

Fig. SOM_5: Change of predicted secondary structure in HSSP homologues vs. random mutations. Native sequences were sampled as subset of our PDB set (main text for more details; note that sampling was imposed by CPU constraints), their homologues were taken from the HSSP database (grey box plots). At a level roughly corresponding to 30% pairwise sequence identity (corresponding to the minimum threshold for sequence homology as defined by HSSP), an average Q3 value of ~63% is reached. This is in contrast to the lower Q3 of ~45% during *in silico* mutation (yellow and green box plots, also s. Fig. 1A, both PAM and Background, main text) at this level of sequence identity. This may be due to the fact that *naturally evolved* homologs underwent natural selection, aiming at the upkeep of already established stable folds resulting in an overall lower rate of change in secondary structure.

Fig. SOM_6:

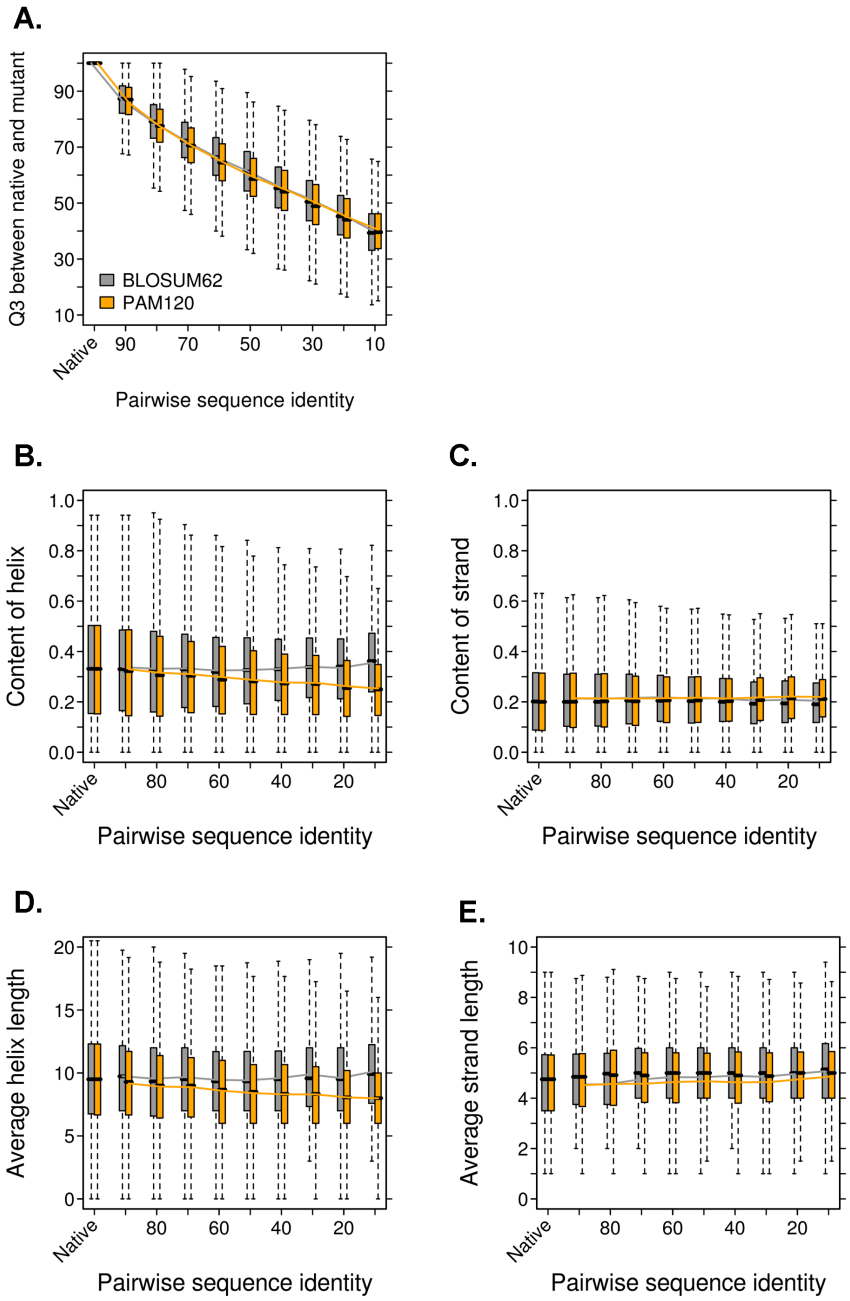

Fig. SOM_6: Secondary structure predictions compared between BLOSUM62 and PAM120 mutations. Q3 levels in both mutation schemes are nearly identical (A). The only major difference is for very high levels of divergence in terms of helices: while the content and length of predicted helices between the native and the final mutant after 69 steps according to BLOSUM are almost identical (B and D), while for PAM120 mutations there appears a slight, albeit statistically insignificant difference, B+D). The overall major result, namely that regular secondary structure is an essential feature of random

sequences is valid equally for mutation according to BLOSUM62 and according to PAM120.

Fig. SOM_7:

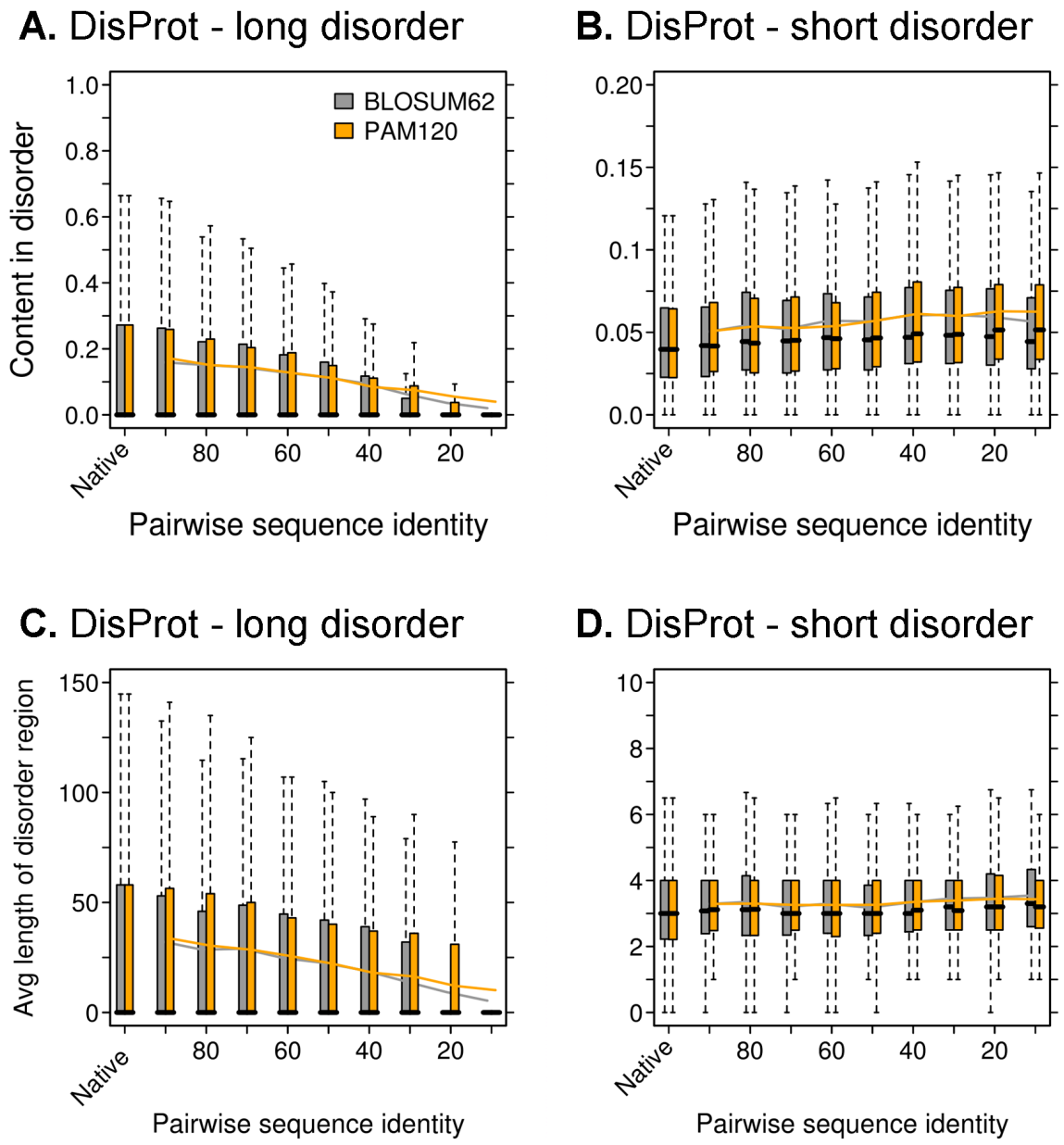

Fig. SOM_7: Comparison of behavior of disorder in DisProt proteins during BLOSUM62 and PAM120 mutation. Overall, both BLOSUM62 and PAM120 mutations seem to behave equally for the features of interest, i.e. content of long (A) and short (B) disorder as well as the lengths of long (C) and short (D) disordered regions. BLOSUM62 seems to work slightly against the upkeep of long and short disordered regions when compared to PAM120.

Fig. SOM_8:

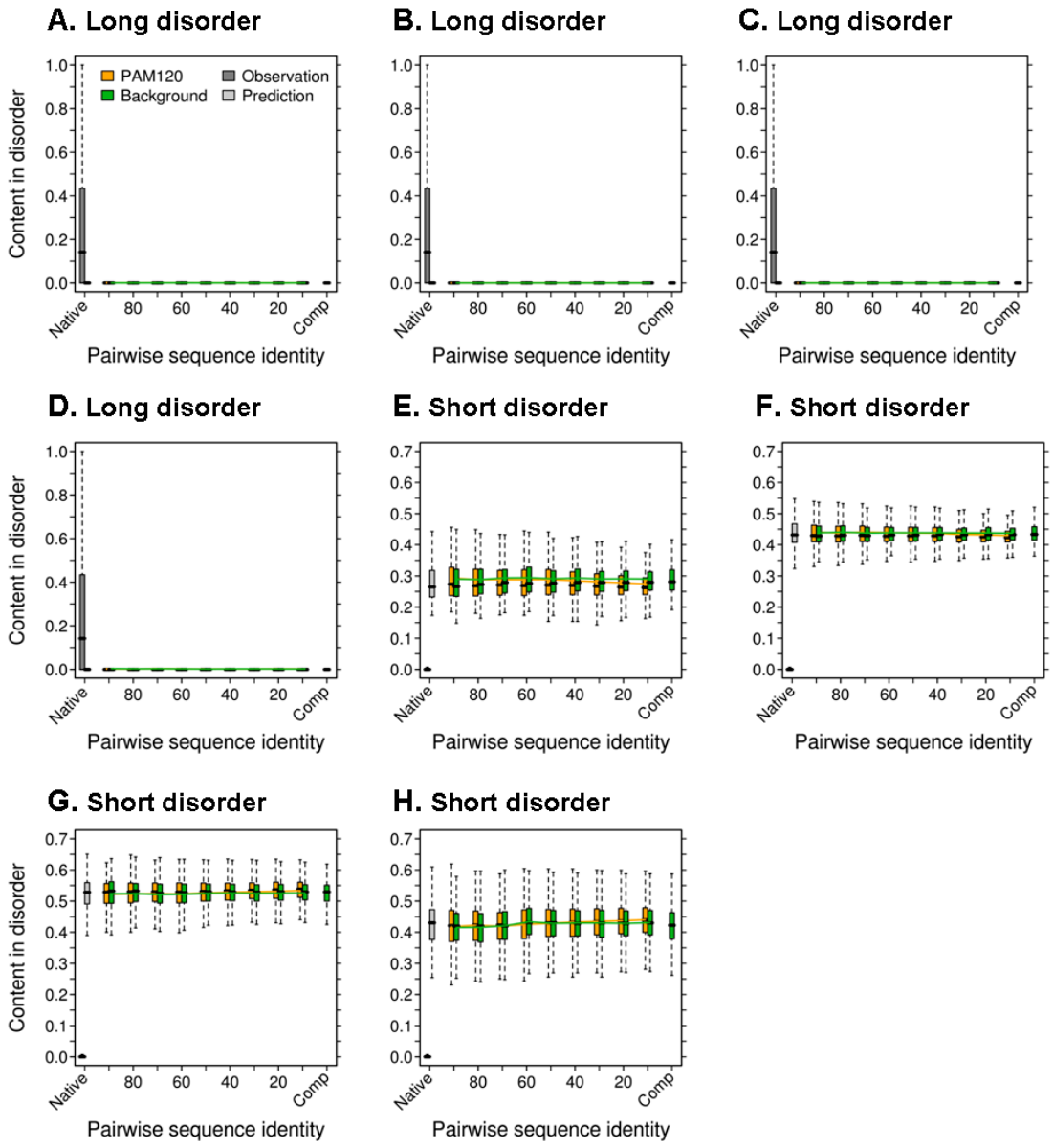

Fig. SOM_8: Simulation of errors in disorder predictions. Errors were inserted randomly into predictions (IUPred) of native sequences of disordered proteins to demonstrate the impact of false predictions. An error constitutes a switch from predicted disorder to order state and vice versa on a per-residue basis. The amount of errors was gradually increased beginning with 20% to 80% (steps of 20%) of residues. Obviously, long disorder (i.e. ≥30 consecutive residues) nearly completely disappears (A-D), since hitting a (disordered) residue in such a region (switching it to ordered state, and thus

disrupting the long region into shorter regions) is very likely using an error rate of 20% (A) or more (B-D). On the other hand, using this error scheme, short disorder (i.e. ≤8 consecutive residues) gets enriched (E-G, 20%-60%) due to the disruption of long disorder and the (undirected) insertion of new disordered residues. Reaching 80% error rate (H), the overall level decreases again, since a situation is approached where the predictions are completely reversed (at 100% error rate). Although very simple, this error scheme shows that even with high error rates the basic trend of upkeep in short disorder during in-silico mutation stays.

Fig. SOM_9:

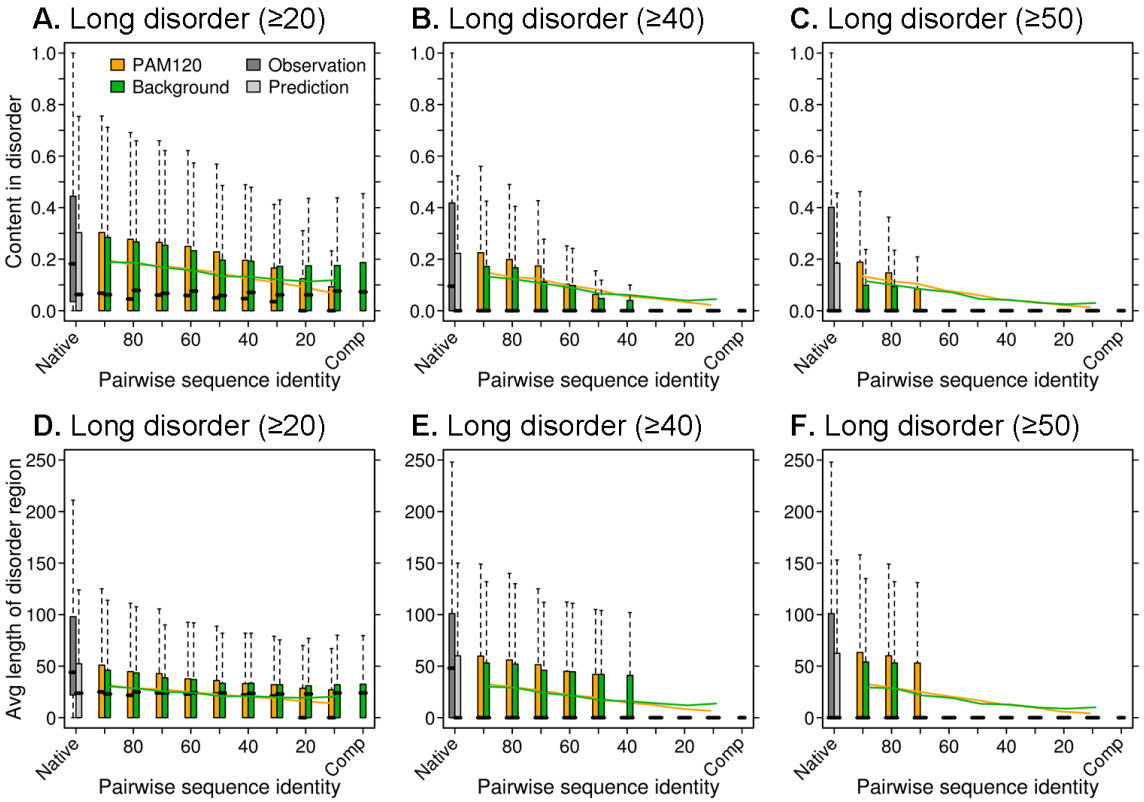

Fig. SOM_9: Different thresholds for definition of long disorder. Coloring and labels as in Fig. SOM_01. Long disorder was investigated in disordered proteins (DisProt dataset) for three different cutoffs (≥20, 40, 50 consecutive residues). Disorder predictions taken from IUPred. The general behavior of disorder content (A-C) and length (D-F) during in-silico mutation seems to be independent of the here chosen thresholds: Both features obviously get reduced during both mutation schemes (see Methods), while this trend is clearer for higher thresholds (B,C for content; E,F for length) where long disorder disappears nearly completely.

Fig. SOM_10:

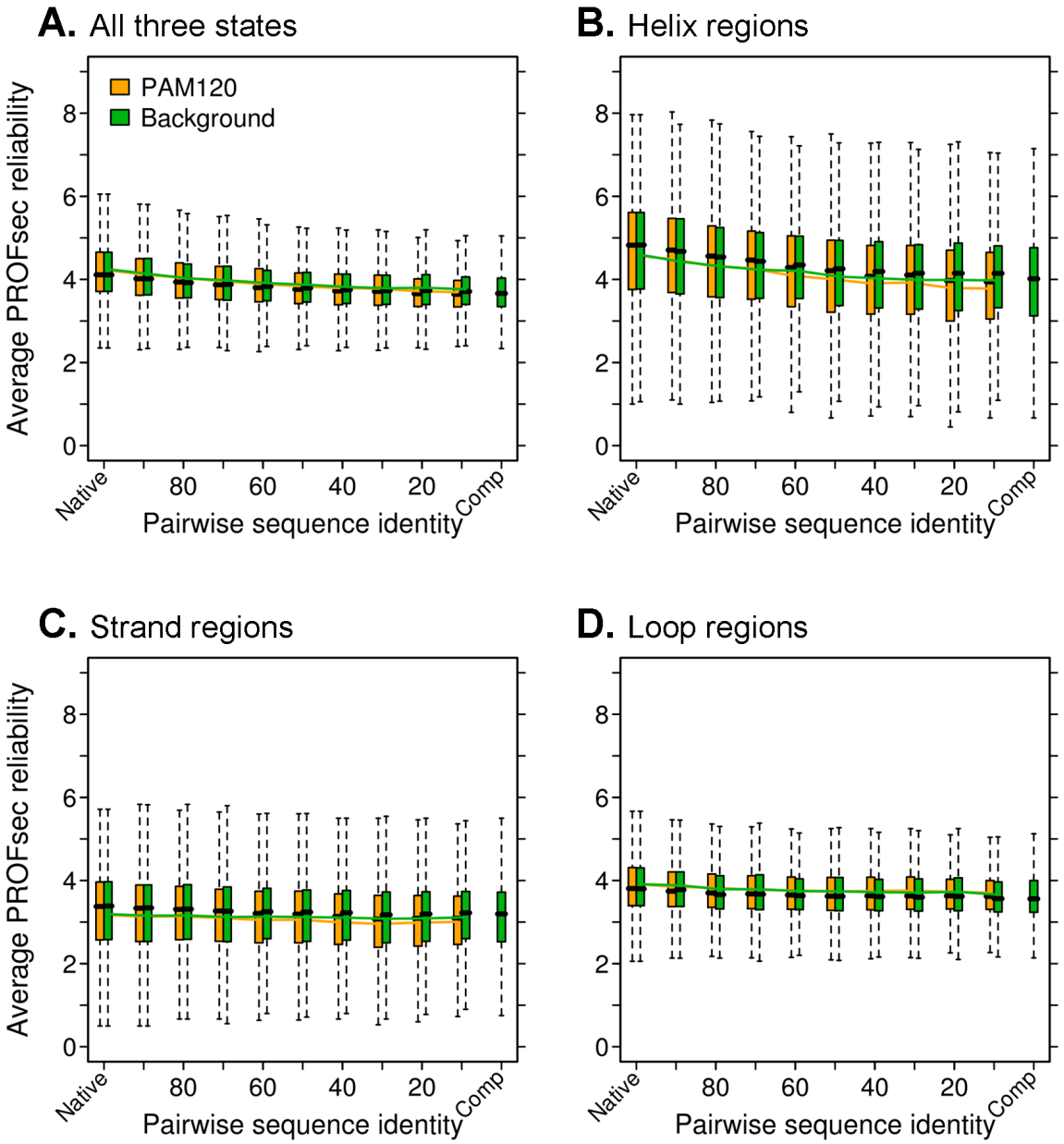

Fig. SOM_10: PROFsec reliability relative to level of divergence. PROFsec's reliability index (RI, values 0-9) indicates the safety of its three-state prediction (helix, strand, loop) on a per-residue basis (0: lowest, 9 highest safety), and is investigated here on different levels of divergence from the native state. The overall (all three states taken together) RI (A) seems to be relatively constant, which is also true for strand (C) and loop (D) regions, while a slightly more dominant decrease for helix regions (B) is visible (mean decrease from RI ~5 to ~4). Note that all RI levels are relatively low due to the circumstance of not taking sequence profiles into account for predictions (see methods). Overall it could be stated that PROFsec predictions do not loose their reliability significantly while moving away from the native state into the realm of artificially created sequences.

Fig.SOM_11:

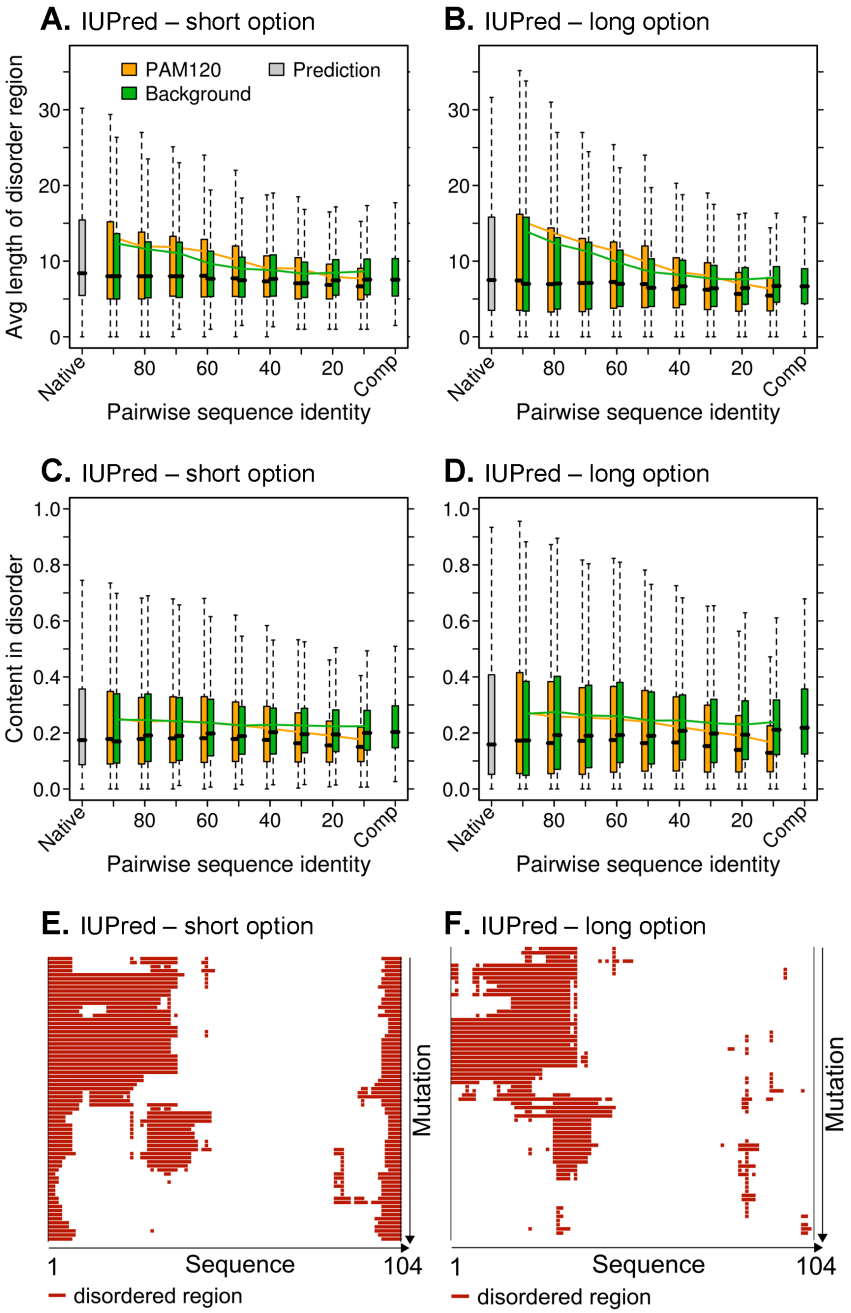

Fig. SOM_11: Behavior of unfiltered disorder during mutation. Disorder predictions are here considered in their raw aspect, i.e. without applying any cutoff to filter out short/long disordered regions. Hence, we present results of IUPred predictions with either short (A,C,E) or long (B,D,F) option being applied. The average length of disordered regions (A,B) and the content in disorder (C,D) get reduced for both mutation schemes

(PAM120: yellow box plots, Background: green box plots) and options, while this trend is more dominant for length compared to content. Since both features stay quite constant for very short regions (s. Fig. 3D, SOM_2A), this suggests that the here observed effect of decrease is mainly caused by longer regions. To illustrate the change of disorder, the mutation trajectories for a representative example (DisProt identifier: DP 00006) are shown for both IUPred options (E short, F long).

Fig.SOM_12:

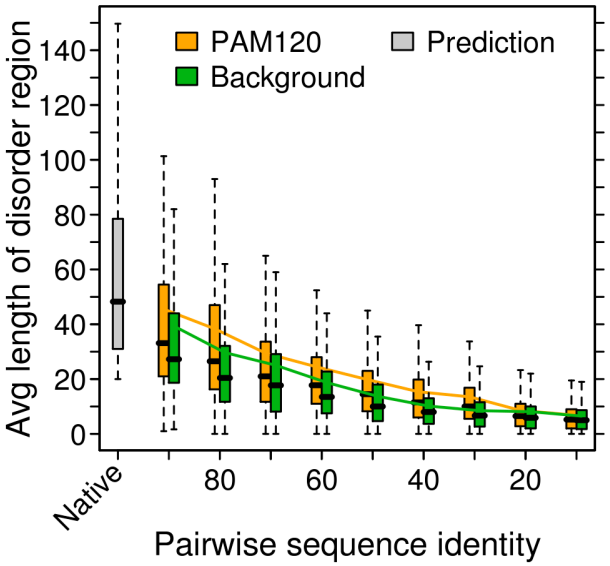

Fig. SOM_12: Decomposition of native long disordered regions. The average length of disorder (without considering any length cutoff) which originates from natively long disordered regions (≥20 consecutive residues) gets clearly reduced during in-silico mutation (PAM120 yellow, Background green boxplots). Regions in native sequences show a mean length of ~50-60 consecutive residues. The length of regions that result from decay of the original one drops down to ~5 residues at very high divergence. This sheds light on behavior that is else obfuscated by the application of a minimum cutoff for the definition of long disorder.

BIOINFORMATICS APPLICATIONS NOTE

Vol. 28 no. 4 2012, pages 601–602
doi:10.1093/bioinformatics/btr705

Databases and ontologies

Advance Access publication December 30, 2011

SNPdbe: constructing an nsSNP functional impacts database

Christian Schaefer[1,2,*], Alice Meier[1], Burkhard Rost[1,2] and Yana Bromberg[3]

[1]Technische Universitaet Muenchen, Bioinformatics – I12, Informatik, Boltzmannstrasse 3, [2]Technische Universitaet Muenchen Graduate School of Information Science in Health (GSISH), Boltzmannstrasse 11, 85748 Garching, Germany and [3]Department of Biochemistry and Microbiology, School of Environmental and Biological Sciences, Rutgers University, New Brunswick, NJ 08901, USA

Associate Editor: Jonathan Wren

ABSTRACT

Summary: Many existing databases annotate experimentally characterized single nucleotide polymorphisms (SNPs). Each non-synonymous SNP (nsSNP) changes one amino acid in the gene product (single amino acid substitution;SAAS). This change can either affect protein function or be neutral in that respect. Most polymorphisms lack experimental annotation of their functional impact. Here, we introduce SNPdbe—SNP database of effects, with predictions of computationally annotated functional impacts of SNPs. Database entries represent nsSNPs in dbSNP and 1000 Genomes collection, as well as variants from UniProt and PMD. SAASs come from >2600 organisms; 'human' being the most prevalent. The impact of each SAAS on protein function is predicted using the SNAP and SIFT algorithms and augmented with experimentally derived function/structure information and disease associations from PMD, OMIM and UniProt. SNPdbe is consistently updated and easily augmented with new sources of information. The database is available as an MySQL dump and via a web front end that allows searches with any combination of organism names, sequences and mutation IDs.

Availability: http://www.rostlab.org/services/snpdbe

Contact: schaefer@rostlab.org; snpdbe@rostlab.org

Received on September 9, 2011; revised on November 11, 2011; accepted on December 17, 2011

1 INTRODUCTION

Resources like dbSNP (Sherry *et al.*, 2001) and UniProt (Bairoch *et al.*, 2005) contain many experimentally determined nsSNPs, but few of these are annotated with respect to function. Some databases [e.g. PMD (Kawabata *et al.*, 1999)] contain experimental annotations of functional effects of mutants. However, these are sparsely populated and do not directly link to dbSNP or UniProt. For the vast majority of mutations lacking experimental annotation, we can gauge functional impact only via *in silico* analysis.

Proper use of computational methods requires specific skills and resources generally inaccessible to medical researchers or experimental biologists. To help, we created an MySQL database readily usable by non-experts. We collected SAASs from PMD, dbSNP, 1000 Genomes (1000_Genomes_Project_Consortium, 2010) and UniProt 'variant's and 'mutant's. We also store 'conflict' records to illustrate how sequencing discrepancies may lead to differing interpretations of the functional significance of a given sequence position. For each SAAS we predict the functional effect using SNAP (Bromberg and Rost, 2007) and SIFT (Ng and Henikoff, 2001). Where available, predictions are augmented by experimental annotations and associated human diseases. We also compute evolutionary conservation of the mutant positions. A web interface provides convenient access to underlying data via organism, sequence and mutation ID queries.

*To whom correspondence should be addressed.

2 DATA SETUP AND RETRIEVAL

Database: SNPdbe mutation data comes from dbSNP, UniProt, 1KG and PMD (Fig. 1A). UniProt and PMD store protein sequences explicitly, while dbSNP links to RefSeq (Pruitt *et al.*, 2007). dbSNP collects 1KG variants with a time delay, so for SNPdbe we mapped all 1KG nsSNPs to RefSeq using Annovar (Wang *et al.*, 2010). We keep only one version of redundant protein sequences, referenced by md5 checksums irrespective of origin. Redundancy is assessed at full-sequence identity (maximum one substitution per sequence) over the entire sequence (+/− leading Met residue). This allows correlating mutations from different sources referencing the same sequence. We currently store 1 362 793 unique SAASs in 158 004 proteins from 2684 organisms covering all kingdoms of life; the top five contributors are human, mouse, rice, cow and rat. For each SAAS we provide the following information: (i) SNAP and SIFT binary predictions of functional effects (neutral/non-neutral). (ii) Evolutionary conservation information from PSIC (Sunyaev *et al.*, 1999), PSI-Blast (Altschul *et al.*, 1997) PSSMs and frequency scores from runs against PDB (Berman *et al.*, 2000) and UniProt. (iii) Functional effects from PMD and UniProt. For human SAASs, disease associations are also available from PMD, UniProt and OMIM (Amberger *et al.*, 2009) (Fig. 1B). (iv) dbSNP evidence and average heterozygosity, and (iv) interesting functional/structural features (UniProt) at the mutation site. Data are stored in an MySQL database and are downloadable as a dump file.

Web interface: The database is web-accessible allowing gene/protein ID/name, disease, sequence (or its md5 hash) and mutant-based queries. Some queries (e.g. md5, gene ID) are exact. Sequence queries are BLAST similarity based. Keyword searches (e.g. disease) are 'loose', i.e. matched to corresponding free text fields. The results page lists all SAASs found within the specified sequence and their functional effect predictions, wild-type/mutant conservation scores, information on disease (human only), experimentally derived functional/structural consequences, changes in position biochemical properties, per-variant validation status and average heterozygosity. This information is also

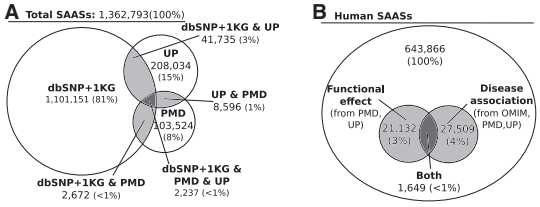

Fig. 1. Venn diagrams describing the overlap of (**A**) all SNPdbe component databases and (**B**) functional and disease annotations of human SAASs. Note that <1% of human SAASs have both functional effect and disease annotations.

accessible via single/batch mutation queries with dbSNP rsids, PMD or SwissVar IDs or SAASs in the X*pos*Y format (and associated sequence). The user can (i) restrict queries to specific organisms or protein keywords; and (ii) search for mutants in similar sequences. Query results may be sorted by different attributes and downloaded in CSV format. Linkouts to referenced web resources are available.

Example: dbSNP rsid 104894374 describes the mutation R157W in the RDH5 gene. This mutation is associated with eye disease, *Fundus albipunctatus* (OMIM 601617.0008). Both SNAP and SIFT predict this substitution to be non-neutral. Indeed, it results in loss of activity in the gene product (PMD A010122). By combining mutation disease associations and their functional effects new inferences can be made about molecular functions altered in disease.

3 CONCLUSION

SNPdbe is designed to fill the annotation gap left by the high cost of experimental testing for functional significance of protein variants. It joins related bits of knowledge, currently distributed throughout various databases, into a consistent, easily accessible and updatable resource. The major features distinguishing SNPdbe from other databases are: (i) the inclusion of a much wider array of organisms and data sources; and (ii) the explicit differentiation between functional/structural effects and disease associations. Furthermore, unlike SNPdbe, existing resources (i) lack experimental annotation of functional/structural changes or offer only single tool (e.g. SIFT) predictions (Mooney and Altman, 2003; Thorn *et al.*, 2010), (ii) are limited to naturally occurring variants (Chelala *et al.*, 2009), (iii) are not consistently updated (Jegga *et al.*, 2007; Wang *et al.*, 2006) or (iv) do not offer pre-computed effects on a large scale (Reva *et al.*, 2011; Wang *et al.*, 2010). SNPdbe's database schema and management scripts are designed to easily handle the addition of new sequences and SAASs and the integration of new predictors and sources of experimental data. Monthly updates are planned. Information about current versions of included databases and statistics is available from SNPdbe website. Our ultimate goal is to make SNPdbe a toolbox for biologists and medical researchers dealing with mutation data. Computationally acquired predictions and annotations found in SNPdbe will help design and prioritize further experimental research.

ACKNOWLEDGEMENTS

We thank to Laszlo Kajan, Guy Yachdav and Tim Karl (TUM) for maintenance of our compute cluster and to those who deposit their experimental data in public databases.

Funding: Alexander von Humboldt Foundation (to C.S., A.M. and B.R.); Rutgers, New Brunswick, start-up funds (to Y.B.).

Conflict of Interest: none declared.

REFERENCES

1000_Genomes_Project_Consortium (2010) A map of human genome variation from population-scale sequencing. *Nature*, **467**, 1061–1073.

Altschul,S.F. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

Amberger,J. *et al.* (2009) McKusick's Online Mendelian Inheritance in Man (OMIM). *Nucleic Acids Res.*, **37**, D793–D796.

Bairoch,A. *et al.* (2005) The Universal Protein Resource (UniProt). *Nucleic Acids Res.*, **33**, D154–D159.

Berman,H. *et al.* (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.

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

Chelala,C. *et al.* (2009) SNPnexus: a web database for functional annotation of newly discovered and public domain single nucleotide polymorphisms. *Bioinformatics*, **25**, 655–661.

Jegga,A.G. *et al.* (2007) PolyDoms: a whole genome database for the identification of non-synonymous coding SNPs with the potential to impact disease. *Nucleic Acids Res.*, **35**, D700–D706.

Kawabata,T. *et al.* (1999) The Protein Mutant Database. *Nucleic Acids Res.*, **27**, 355–357.

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

Ng,P.C. and Henikoff,S. (2001) Predicting deleterious amino acid substitutions. *Genome Res.*, **11**, 863–874.

Pruitt,K.D. *et al.* (2007) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.*, **35**, D61–D65.

Reva,B. *et al.* (2011) Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic Acids Res.*, **39**, e118.

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

Sunyaev,S.R. *et al.* (1999) PSIC: profile extraction from sequence alignments with position-specific counts of independent observations. *Protein Eng.*, **12**, 387–394.

Thorn,C.F. *et al.* (2010) Pharmacogenomics and bioinformatics: PharmGKB. *Pharmacogenomics*, **11**, 501–505.

Wang,K. *et al.* (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.*, **38**, e164.

Wang,P. *et al.* (2006) SNP Function Portal: a web database for exploring the function implication of SNP alleles. *Bioinformatics*, **22**, e523–e529.

Schaefer and Rost *BMC Genomics* 2012, **13**(Suppl 4):S4
http://www.biomedcentral.com/1471-2164/13/S4/S4

BMC Genomics

PROCEEDINGS Open Access

Predict impact of single amino acid change upon protein structure

Christian Schaefer[1,2*], Burkhard Rost[1,2,3,4,5]

From SNP-SIG 2011: Identification and annotation of SNPs in the context of structure, function and disease
Vienna, Austria. 15 July 2011

Abstract

Background: Amino acid point mutations (nsSNPs) may change protein structure and function. However, no method directly predicts the impact of mutations on structure. Here, we compare pairs of pentamers (five consecutive residues) that locally change protein three-dimensional structure (3D, RMSD>0.4Å) to those that do not alter structure (RMSD<0.2Å). Mutations that alter structure locally can be distinguished from those that do not through a machine-learning (logistic regression) method.

Results: The method achieved a rather high overall performance (AUC>0.79, two-state accuracy >72%). This discriminative power was particularly unexpected given the enormous structural variability of pentamers. Mutants for which our method predicted a change of structure were also enriched in terms of disrupting stability and function. Although distinguishing change and no change in structure, the new method overall failed to distinguish between mutants with and without effect on stability or function.

Conclusions: Local structural change can be predicted. Future work will have to establish how useful this new perspective on predicting the effect of nsSNPs will be in combination with other methods.

Background

Protein structures very robust under sequence change

Evolution creates the specific protein landscape that we observe today. Mutations are random but selection is the driving force that shapes the observable protein variety by favoring those deviations that maintain or improve phenotype. This constrained sampling process explains the sequence diversity compatible with a given protein three-dimensional (3D) structure: over 50-80% of all residues can be changed without altering structure significantly [1-3].

Local structure change can impact phenotype

Although many different sequences map to similar structures, point mutants can change structure dramatically [4-6]. Some of the intricate details of 3D structures are crucial for function. Therefore, such local conformational changes may impact protein function and may cause disease. Usually, this is more likely for structure changes connected to binding sites. For instance, the disruption of hydrophobic interactions, or the introduction of charged residues into buried sites, or mutations that break beta-sheets often impact phenotype severely and raise the susceptibility for disease [7-9]. Using 83 X-ray mutant structures from 13 classes of proteins, an early work pioneered the prediction of local structural changes by expert rules operating on position-dependent rotamers [10]. It is unclear, how well such an approach would cope with the protein variety found in the current PDB [11]. Thus, we followed a different approach. We compiled a set of structurally superimposed pairs of protein fragments with identical sequence except for one central residue mismatch, and applied machine-learning to predict structural change from sequence.

Methods

Central pentamer data

We extracted 146,296 protein chains from X-Ray structures in the Protein Data Bank (PDB, July 2010) [11].

* Correspondence: schaefer@rostlab.org
[1]TUM, Bioinformatics - I12, Informatik, Boltzmannstr. 3, 85748 Garching, Germany
Full list of author information is available at the end of the article

Then we applied two techniques for redundancy reduction. The first set (dubbed "cdhit98") contained 24,890 chains; it resulted from clustering with CD-HIT [12] to a level at which no pair had over 98% percentage sequence identity. The second set (dubbed "hval0") contained 3,767 chains; it resulted from filtering at HVAL>0 [2,3,13] (corresponding to ~20% maximal pairwise sequence identity for alignments over 250 residues). We chopped each chain in each set into all overlapping fragments of five consecutive residues (pentamers), removing: (i) pentamers with chain breaks (peptide bond length >2.5Å, as defined in DSSP [14]), (ii) pentamers with non-standard amino acids, and (iii) all but the first set of atomic coordinates for residues with alternative locations. Each pentamer from the first set (cdhit98) was paired with each pentamer from the second set (hval0).

We selected pairs of pentamers that differed only in the central amino acid, and that originated from proteins with over 30% overall percentage pairwise sequence identity. We also filtered out pairs for which either fragment was already in a much larger fragment that fulfilled the above criteria. This procedure yielded 35,533 pentamer pairs. For each pair, we calculated the root mean square displacement (RMSD) over all C-alpha atoms after optimal superposition of the two pentamer backbones (McLachlan algorithm [15] as implemented in ProFit [16]). To turn the continuous RMSD differences into a binary problem (mutant changes structure or not), we had to decide what constitutes a structural effect and what is neutral in that sense. In lack of a scientifically meaningful definition for structural change of pentamers, we chose thresholds that appeared reasonable given the observed distributions and that separated all pentamer pairs into an even amount of structurally neutrals and non-neutrals. We defined RMSD values <0.2Å as structurally neutral and values >0.4Å as structurally non-neutral, i.e. as structural change; we ignored all pairs in between these two. These particular thresholds assigned 12,046 pentamer pairs to the class of "structural change" and 13,675 to the class "neutral". For each such pair we randomly designated one fragment as wild type fragment and the central mismatch residue of the other fragment as the mutant amino acid.

Additional functional data

For comparison, we also used two data sets that had been used previously (Additional file 1). The first set comprised 12,461 functionally neutral and 35,585 functional effect mutants from 3,444 proteins [17,18]. The second consisted of 657 mutants having an effect on protein stability and 652 mutants with no effect on stability covered by 47 proteins [19,20]. Mutations leading to a change in the Gibbs free energy ($\Delta\Delta G$) < -1 kcal/mol or >1 kcal/mol were considered as non-neutral (i.e. both stabilizing and destabilizing mutations were taken as assays of change); all other mutations were treated as neutral (i.e. no effect).

Additional prediction methods

Various methods predict other aspects of the impact for amino acid changes, e.g. effects on protein function or stability. In particular, we applied SNAP [17] and I-Mutant3 [21] to test their discriminative power on our data sets. Both methods return raw numerical scores reflecting direction and reliability of the prediction. SNAP values range from -100 (neutral for function) to 100 (change of function). The distance of the actual prediction to the decision boundary (0) reflects the reliability of the prediction and the severity of the predicted effect (large distance = high reliability and severity [17]). I-Mutant3 predicts the $\Delta\Delta G$ value upon mutation. We adhered to the same decision cutoffs as mentioned above to define neutral and non-neutral.

Prediction method: basics

We applied logistic regression to learn the structural change upon amino acid change. Logistic regression is a parameter-free machine-learning algorithm; we adhered to an implementation offered by the LIBLINEAR package (L2-regularized logistic regression, dual) [22].

Many protein features may be relevant for the given prediction task. Our feature construction procedure adhered to a protocol established during the development of SNAP [17]. All features were derived from protein sequence alone and were extracted from PredictProtein [23], a wrapper that combines a large number of independent prediction methods. We used three conceptually different types of features: (1) global features describing the global characteristics of a protein, (2) local features describing one particular pentamer and its immediate sequence neighborhood, and (3) difference features that explicitly describe sequence-derived aspects by which wild type and mutant amino acid differ.

(*1*) *Global features:* We represented sequence length as four different values each representing a length interval (1-60, 61-120, 121-180, 181-240 consecutive residues). The bin that represented the sequence length was set to 0.5, bins below were assigned to 1, bins above to 0. Amino acid composition was encoded by 20 values representing relative frequencies of standard amino acids. We predicted secondary structure and solvent accessibility using PROFphd [24,25]. Three values represented the relative content of residues in predicted helix, strand and loop conformation and, similarly, three values were used to encode the relative content of predicted buried, intermediate and exposed residues.

(*2*) *Local features:* We used features that described the local sequence neighborhood of the amino acid change. We considered window lengths of 1 (position of change

only), 5, 9, 13, 17 and 21 consecutive residues centered on the position of change. Values were normalized to the interval [0, 1]. The biochemical characteristics of an amino acid influence the local structural conformation. We considered six different structural and biochemical propensities: mass, volume [26], hydrophobicity [27], C-beta branching [28], helix breaker (only proline) and electric charge of side chain. Evolutionary information contained in sequence profiles is a valuable source to obtain knowledge about which amino acids are compatible with a specific region in the protein. While some residues are tolerated others could disrupt structure. We used position specific scoring matrices (PSSMs), relative amino acid frequencies and the information content per alignment position taken from PSI-BLAST [29] runs (options: -j 3 –b 3000 –e 1 –h 1e-3) against a sequence database consisting of UniProt [30] and PDB [11]. Sequences were redundancy-reduced to a level where no protein pair had more than 80% sequence identity [12]. Furthermore, we took position-specific independent counts (PSIC [31]) and adhered to a protocol necessary for sequence extraction and generation of multiple alignment as described elsewhere [17]. In addition, we used the following predicted structural and functional features: secondary structure [32,33] and solvent accessibility [24,25,32], protein flexibility [34], protein disorder [35-38], protein-protein interaction hotspots [39-41] and DNA-binding residues [42]. Most prediction methods used to generate features returned both a discrete prediction and a score reflecting the strength and reliability of the prediction. We incorporated both outputs in our feature set. Two-state predictions (disorder, protein and DNA interaction) were encoded as two mutually exclusive combinations of 1 and 0, each representing the presence (1) and absence (0) of a state (e.g. disorder vs. no disorder). Three-state predictions (secondary structure elements helix, strand, other and solvent accessibility states buried, intermediate, exposed) were handled similarly. Flexibility was predicted as a numerical value only. We considered information about the location of the site of change in the sequence relative to a protein domain as an important feature. For example, a hydrophobic-to-polar exchange within the core of a domain may have a more severe impact on local structure than a change that happens in a surface loop. We extracted relevant per-residue information out of the protein family database Pfam-A [43] using the output from HMMER3 [44]. Of specific interest was the information about whether the residue resided in a domain, the conservation of that position within the domain alignment, how well the residue fitted into the alignment position and the posterior probability of that match.

(*3*) *Difference features*: Of particular interest were features that captured the difference in characteristics between the two differing central amino acids in a pair of pentamers. We represented the difference of a particular property separately by its absolute and its sign, encoded as 0 (negative) or 1 (positive). The following properties were encoded in that respect: Change in any of the six amino acid propensities, difference in conservation scores (PSSM, relative frequency, PSIC), change in IUPred predictions for both short and long disorder, change in predicted secondary structure and solvent accessibility. For the latter two we ran PROFphd on raw sequence rather than sequence profile. Although this mode resulted in reduced prediction performance, it allowed us to observe an actual difference in the prediction outcome, which would have been disguised by the use of sequence alignments otherwise.

Prediction method: feature selection

We concentrated the training of our model only on the most predictive sequence features. Toward this end, we considered one fifth of the pentamer pairs (2,243 structurally non-neutral, 2,882 neutral) and ensured that those pairs were derived from proteins without significant sequence similarity (EVAL>10-3) to any protein in the remaining four fifth of the data. Those 5,125 instances were further partitioned into ten subsets. Nine such sets participated in training a logistic regression model, while its performance was tested on the remainder. We rotated ten times over all sets such that each instance served once during testing and training and guaranteed that no significant sequence similarity existed between train and test folds (EVAL>10-3). Before each new rotation, a set of features for training and testing the model was determined by the following iterative protocol. We started with one feature and established its predictive performance during one complete rotation as explained above. We did that for all global and difference features as well as every combination between local features and window lengths. We measured feature performance by means of average AUC (area under the receiver-operator curve) derived from rotating ten times over the testing folds. The best performing feature was automatically included for the subsequent evaluation of the remaining features. We stopped this forward selection after no further increase in average AUC>0.001 was observed.

Performance estimates

We assessed performance only on the test sets (as described above). In lack of a biological intuition for how to measure the success of our prediction method, we fell back to standard measures. Following the typical acronyms, we used TP (true positives) to denote pairs correctly predicted to change structure (positive) and FP (false positives) are neutral pairs predicted as change. In analogy, TN (true negatives) describes correctly predicted neutral

pairs (no change) and FN (false negatives) are structure-changing pairs incorrectly predicted as being neutral. With these, we compiled ROC (Receiver Operating Characteristic) plots, as well as the True Positive Rate (TPR), and the corresponding False Positive Rate (FPR) defined by:

$$TPR = TP / (TP + FN) \qquad FPR = FP / (FN + TN) \quad (1)$$

The area under the ROC-curve (AUC) averaged over ten rounds of training and testing served as a single performance estimator. We also employed the overall two-state accuracy, often referred to as the Q_2 measure. Finally, we monitored class-specific values for AccuracyC, i.e. the accuracy for the class "structural change", AccuracyN (accuracy for the class "neutral"), CoverageC (coverage for class "change") and CoverageN (coverage neutral) defined by:

$$Q_2 = (TP + TN) / (TP + FP + TN + FN) \quad (2)$$

$$AccuracyC = TP / (TP + FP) \qquad CoverageC = TP / (TP + FN)$$
$$AccuracyN = TN / (TN + FN) \qquad CoverageN = TN / (TN + FP)$$

Our logistic regression model yielded a probability for an instance to be structurally non-neutral rather than a discrete class label. By iterating over different probability thresholds, we sampled a ROC-like space of Accuracy-Coverage pairs for each of the two classes.

Box plots

We presented distributions through box plots. The lower and upper box edges depict the first and third quartile, respectively. The length of a box is the interquartile range of the distribution. The bold bar inside the box represents the median, while dashed lines reach to the most extreme data point that is no more than 1.5 times the interquartile range away from the upper or lower box edge. It is worth noticing that per definition the box covers half of the distribution.

Results and discussion

Fitting parameters to observations easily ends in the trap of over-optimization [45]. We have addressed this issue in two ways (Methods). Firstly, we carefully applied standard cross-validation techniques. This included setting pentamer pairs aside that were used only for feature selection, ascertaining minimal sequence similarity between cross-validation sets, and avoiding to over-sample the data set. Secondly, we compared the final method on completely different data sets.

Evolutionary and structural features most predictive

Our forward selection scheme (Methods) yielded the following features as most informative (Fig. 1): difference in PSIC between "native" and "mutant", predicted secondary structure (w=17), BLAST information for each residue (w=21), residue flexibility (w=21), difference in PSSM and predicted secondary structure between "native" and "mutant", HMMER scores for fitting amino acids into a PFam domain alignment (w=13), predicted protein-protein interaction hotspots (w=13), and finally the amino acid volume (w=5). Due to the specific encoding of those properties (Methods), the overall feature space covered 147 numerical feature values.

Three features dominate, most features unstable

For the final assessment of our method, we applied full cross-validation. However, in this paragraph, focus is on assessing the relative contribution of input features. Toward this end, we only used one fifth of the data as one attempt to avoid over-fitting. The numbers are, therefore, only relevant in a relative way.

The success of the method was dominated by the first three features, as indicated by the steepest ascent in average AUC (Fig. 1, first three box plots and solid line). Already the very first property alone (difference in PSIC values between wild type and mutant residue) gave an AUC of almost 0.72 (compared to the random value of 0.5). With the third feature (BLAST information per position, w=21), the discrimination reached an AUC of almost 0.82, close to the performance maximum. The inclusion of the last feature (residue volume) gave an AUC of ~0.84 (Fig. 1, last box plot). Thus, the most informative feature increased the AUC by 0.2, the last six together by only one tenth of this.

The per-feature performance varied strongly in their AUC distributions (Fig. 1, long box plots). While this variance was most pronounced for the first feature (PSIC difference), the trend continued throughout the feature selection (decrease in variability easily explained by the decreasing performance). In the performance plateau regime, features were no longer distinguishable by the distributions of their ten AUC values (Fig. 1, nearly complete box plot overlap after the third feature). Nevertheless, we stopped the feature selection when the performance did not improve more than AUC>10-3. This early stop was implemented as another safeguard against over-fitting.

Sequence-based prediction of structural impact successful

All performance measures reported in following were compiled from a 10-fold cross validation (Methods). The logistic regression model estimates the probability for structural change. Through a simple threshold, this probability gives a binary prediction (e.g. change>0.5, neutral≤0.5) with an overall two-state per-residue accuracy Q_2>72%. However, we also established ROC-curves and accuracy-coverage plots by dialing through the whole

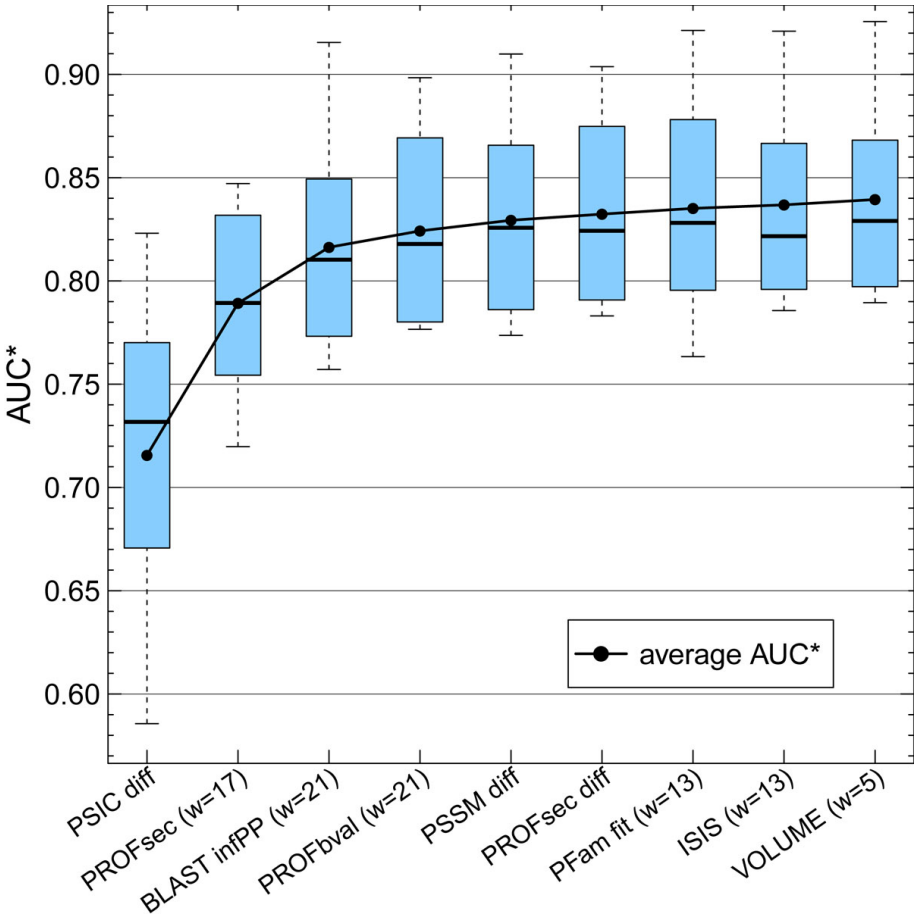

Figure 1 Structural and evolutionary features most predictive. Input features according to their cumulative contribution to performance measured by AUC, i.e. the area under the ROC curve (AUC* indicates that these values refer to results for a subset of the full cross-validation set). Our forward feature selection scheme suggested that three features raised performance above 0.8: evolutionary information (PSIC [31] diff), predicted secondary structure (from PROFsec [32,33]) around mutant (mutant position ± 8, i.e. 17 input units), and the PSI-BLAST information per residue for 21 consecutive residues. Additional six features only marginally increase performance up to mean AUC* ~0.84: predicted flexibility (PROFbval, w=21), difference in both PSI-BLAST PSSM (PSSM diff) and predicted secondary structure scores (PFOFsec diff), the fit of change position into a PFam domain (PFam fit, w=13), scores for predicted protein-protein interaction hotspots (ISIS, w=13) and residue volumes (VOLUME, w=5). High variability in AUC* distributions (long box plots, strong overlap between box plots) indicates instability in selected features.

spectrum of probability values (Fig. 2A). The final model reached an overall AUC of ~0.8.

Both above measures assess overall performance without explicitly revealing per-class (change/neutral) levels. We investigated pairs of coverage/accuracy values sampled at different probability thresholds. More than half of neutral and non-neutral predictions (52%) reached around 80% accuracy (Fig. 2B); for higher accuracy, the correct predictions were dominated by predictions of effect.

These results suggested that sequence suffices to predict the impact of point mutations upon structure through machine learning. This is particularly remarkable in light of the fact that pentamer conformations depend crucially on their structural environment outside the windows that we have considered as input features in our prediction method [46-48].

Structural effect predictions enriched in functional impact

Our explicit objective was to predict the impact of single point mutations upon local structure. The implicit objective was to also develop a new perspective that aids in the prediction of how mutations affect function. While it is clear that the subset of all mutations that locally change structure will be enriched in mutations that also affect function, the inverse is not true: mutations that do not change structure may or may not change function, i.e. will not be enriched in "functionally neutral". If our prediction method captured important aspects of structural change, at best its prediction of

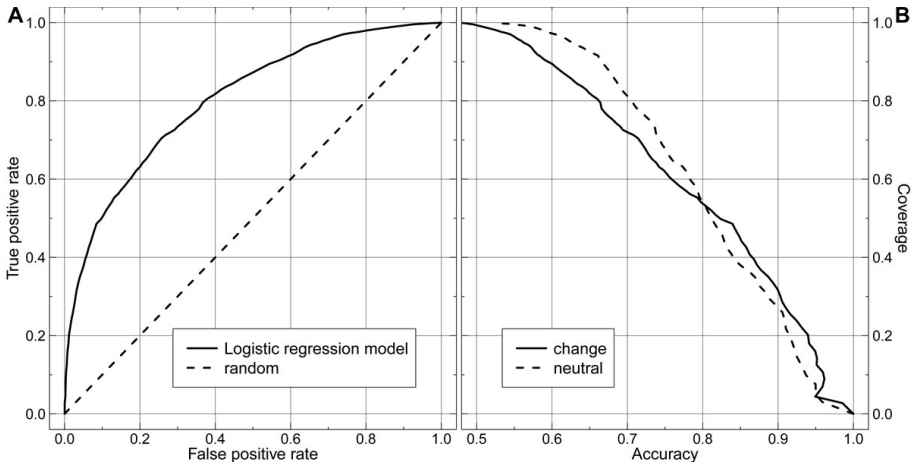

Figure 2 Good discrimination between pentamers with and without effect. All values refer to full-cross validation averages of data not used for feature optimization. Left panel (A): our best model (solid line) reached an AUC of 0.8, compared to random predictions (dashed line) with AUC=0.5. Right panel (B): predictions for effect on structure (change) and predictions for no effect (neutral) reached similar levels on the accuracy vs. coverage plot equation (2).

structural impact will be enriched in those with functional impact.

We tested this alternative perspective on performance in two ways. On the one hand, we used a data set distinguishing amino acid mutations (nsSNPs) that impact function from those that do not. On the other hand, we used a data set of mutants that do and do not impact protein stability. Two results stood out from this analysis. First, mutations predicted to affect structure were enriched in those that also affect function (Fig. 3, ascending dashed curve). Second, the enrichment was proportional to the severity of predicted structural change: starting at over 76% to values over 81% at a probability >0.9 (Fig. 3). We observed a similar trend for the stability data: enrichment in predicted structural effect mutations was 8-13 percentage points above random (random: 50%, enrichment: 58%-63%, Fig. 3). Due to little sample size, the stability enrichment was less significant than that for functional impact.

The above results strongly suggested that our method captured important information beyond its explicit training task. The enrichment over the background might not seem particularly strong (for function: background about 74% vs. 81% predicted, for stability: background 50% vs. 63% predicted). However, it remains unclear what to compare this enrichment with: some mutations affect structure but not function. So what would the enrichment become if we had the complete experimental information correlating all possible assays for structure and function change? Does our method pick up a significant fraction of the possible signal? We have no means of answering this question. However, our prediction method undoubtedly captured a signal pointing into the expected direction: The increasing severity of structural effect upon amino acid change is linked with an accumulation of mutants having an effect on protein function or stability, and this achievement was truly "novel" and it provides information that seems orthogonal to what any other method could have provided.

Signal for the reverse: predicted functional impact more pronounced in structural change

In the previous paragraph, we established that our structure impact predictions capture some signal of functional change. What about the opposite, i.e. to which extent do methods that aim at predicting impact on function (e.g. SNAP [17]) and on stability (e.g. I-Mutant3 [21]) correctly capture the impact of mutations upon structure? First, we provided the "background" by the application of our structural effect method (Fig. 4A+D; data for cross-validation). Both SNAP (Fig. 4B+E) and I-Mutant3 (Fig. 4C+F) failed to separate mutations with and without impact on structure. SNAP at least was able to observe some signal: very few mutations with impact on structure were predicted at scores corresponding to predictions of strong effect upon function. At the default probability threshold of 0.5 our method correctly predicted 69% of all effect (Fig. 4D left dark blue bar), and 76% of all the neutral pentamers (Fig. 4D, right light blue). The corresponding numbers were 39% functional effect in structural effect / 88% functional neutral in neutral for SNAP (Fig. 4E), and 33% effect on stability in structural effect / 72% no effect on stability in neutral for I-Mutant3 (Fig. 4F).

One conclusion from applying SNAP and I-Mutant3 to our data is that only our method succeeded in managing the task that we had set. One possible explanation is that

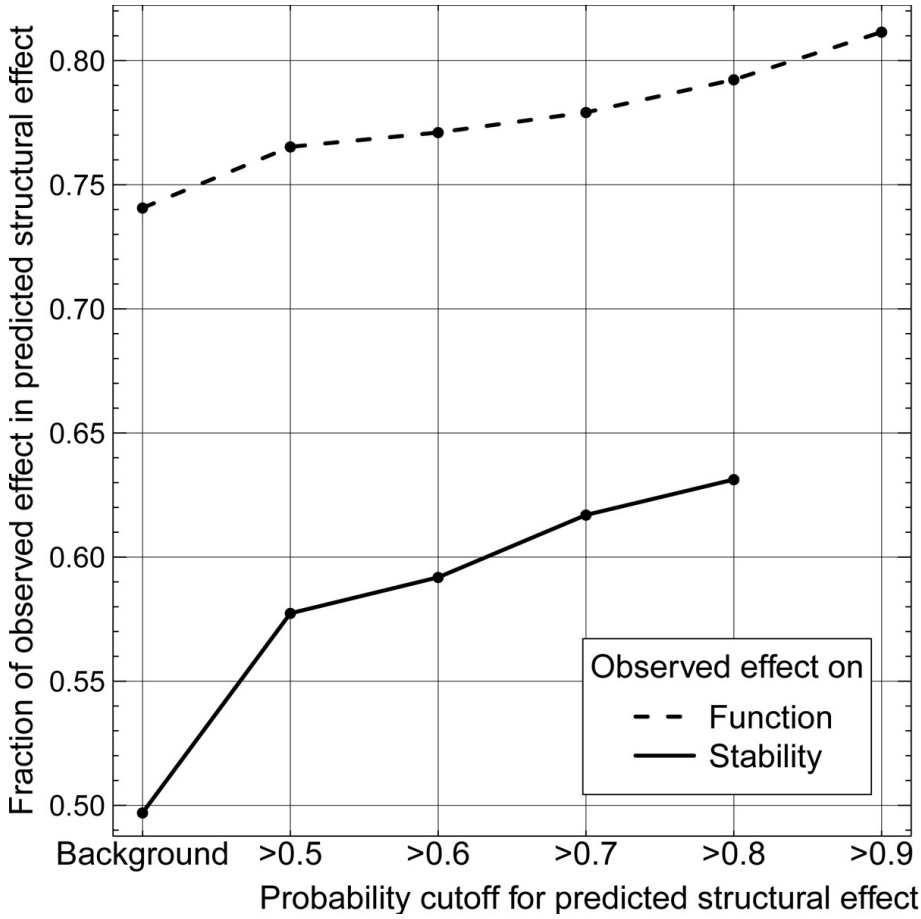

Figure 3 Mutations predicted to affect structure often impact function. We investigated the subset of mutants predicted by our method to change structure from uncertain predictions (>0.5) to very strong predictions (>0.9). Residues in this subset, we found more often than expected to also have an observed effect in other data sets than used for our method, namely on protein stability (solid line) and protein function (dashed line). This suggests that strong structural change upon amino acid change results in increased likelihood to alter function or stability.

our task is incorrectly formulated, i.e. our data set of pentamers with and without local structural change is wrong. Imagine, we assigned labels to pentamers randomly. Then SNAP and I-Mutant3 would fail. If the labels had truly been random, our own method would fail, too. Assume they are not random but biophysically meaningless (e.g. mutations to aromatic amino acids cause change, all others are neutral). If this assumption were fully true, our method would not have picked up a signal in the other data sets that we tested (Fig. 3). Furthermore, if our data set were fully non-sense, SNAP could not have picked up a weak signal. The fact that I-Mutant3 does not pick up a signal may point to the difference between local changes – as targeted here – and global changes – as targeted by I-Mutant3.

All the above considerations support the view that our definition of local structural change captures an important feature of the response of proteins to amino acid changes, and that the method introduced here succeeds at solving the task that we posed.

Conclusions

How do point mutations change the life of a protein? Here, we introduced three new views toward tackling this question. Firstly, we introduced a different perspective of change. Structural effect by our definition is perceived as two protein fragments having a significant dissimilarity in backbone conformation. Secondly, we created a new dataset that allowed us to successfully train a machine-learning model with the incentive to separate structural neutral from non-neutral fragments. Thirdly, we established that both our method and definition of structural change also capture to some extent the impact of change on protein function. It remains to be investigated in more detail how exactly the new method can help in annotating the impact of amino acid changes and nsSNPs.

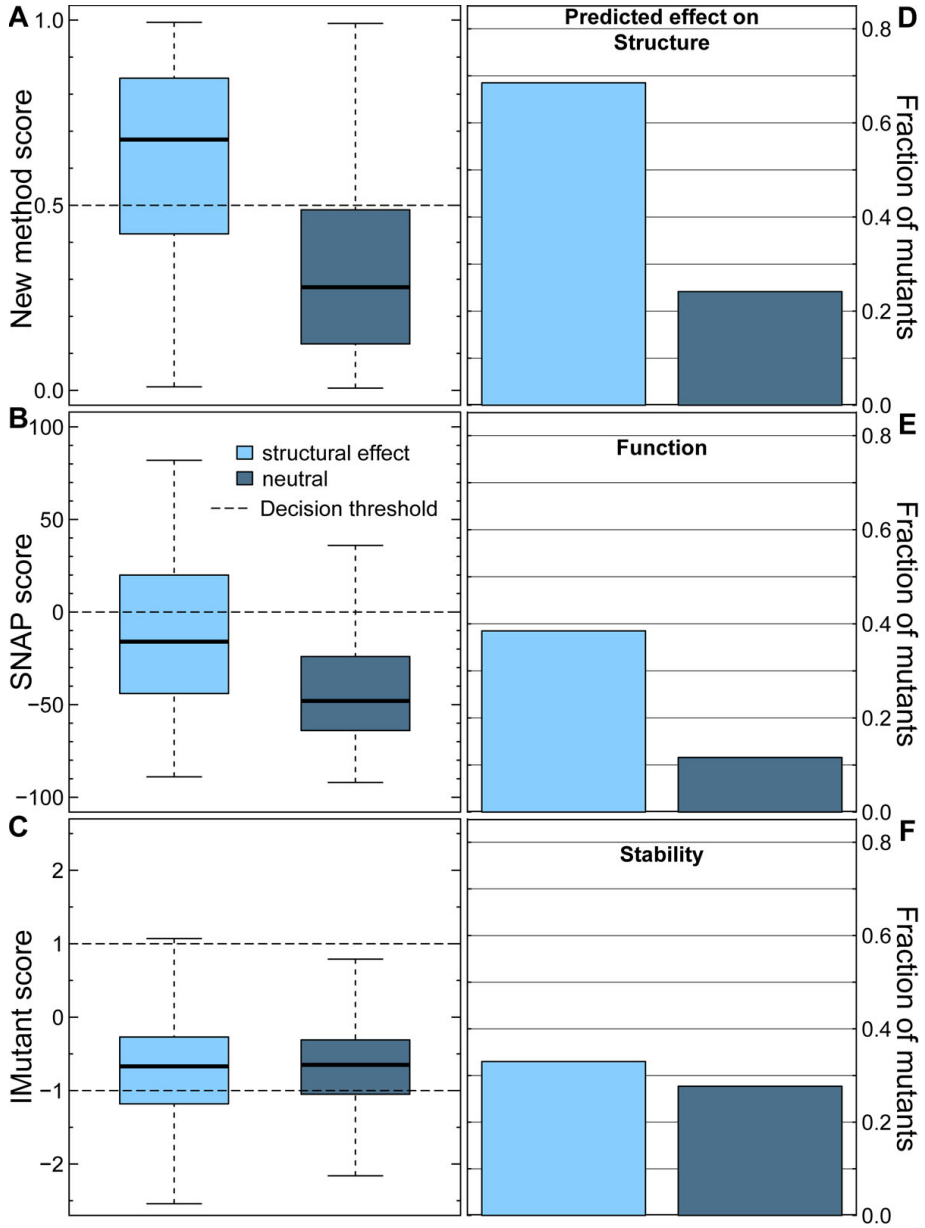

Figure 4 Correlation between structure and function not picked up by other methods. We applied three prediction methods to our dataset of structural effect: (A, D) the new method introduced here, (B, E) SNAP [17] predicting impact on function, and (C, F) I-Mutant3 [21] predicting the impact on stability. In lack of a better alternative, we chose the default threshold for each method (horizontal dashed lines) to distinguish neutral from effect. The method introduced here that is specialized to separate structural effect from neutral performs best at this task (A: little overlap between boxes; note: data in cross-validation mode of our method). The distributions from SNAP (functional effect prediction) and I-Mutant3 (stability prediction) both do not capture the structure signal.

Additional material

Additional file 1: Datasets of mutants with observed effects on function and stability. Archive of the two different mutant sets with observed effects along with predictions of their effect on local structure.

Acknowledgements
We thank Yana Bromberg (Rutgers), Marco Punta (Sanger) and Ulrich Mansmann (LMU Munich) for helpful discussions. Special thanks go to Laszlo Kajan, Guy Yachdav and Tim Karl (TUM Munich) for maintenance of our compute cluster and to Marlena Drabik (TUM Munich) for administrative support. Particular thanks to the anonymous reviewers and the two editors (Emidio Capriotti, University of Balearic Islands, and Yana Bromberg, Rutgers)

for their important help. Last not least, thanks to those who deposit their experimental data in public databases and those maintaining those databases.
Funding: CS and BR were funded by Alexander von Humboldt Foundation. This article has been published as part of *BMC Genomics* Volume 13 Supplement 4, 2012: SNP-SIG 2011: Identification and annotation of SNPs in the context of structure, function and disease. The full contents of the supplement are available online at http://www.biomedcentral.com/bmcgenomics/supplements/13/S4.

Author details
[1]TUM, Bioinformatics - I12, Informatik, Boltzmannstr. 3, 85748 Garching, Germany. [2]TUM Graduate School of Information Science in Health (GSISH), Boltzmannstr. 11, 85748 Garching, Germany. [3]Institute of Advanced Study (IAS), TUM, Boltzmannstr. 3, 85748 Garching, Germany. [4]New York Consortium on Membrane Protein Structure (NYCOMPS), TUM Bioinformatics, Boltzmannstr. 3, 85748 Garching, Germany. [5]Department of Biochemistry and Molecular Biophysics, Columbia University, 701 West, 168th Street, New York, NY 10032, USA.

Authors' contributions
CS carried out the data analysis, programming, and helped to draft the manuscript. BR conceived and supervised the project, and helped to draft the manuscript. All authors read and approved the final manuscript.

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

Published: 18 June 2012

References

1. Shakhnovich EI, Gutin AM: **Influence of point mutations on protein structure: probability of a neutral mutation.** *Journal of theoretical biology* 1991, **149(4)**:537-546.
2. Sander C, Schneider R: **Database of homology-derived protein structures and the structural meaning of sequence alignment.** *Proteins* 1991, **9(1)**:56-68.
3. Rost B: **Twilight zone of protein sequence alignments.** *Protein engineering* 1999, **12(2)**:85-94.
4. Eriksson AE, Baase WA, Zhang XJ, Heinz DW, Blaber M, Baldwin EP, Matthews BW: **Response of a protein structure to cavity-creating mutations and its relation to the hydrophobic effect.** *Science* 1992, **255(5041)**:178-183.
5. Garcia-Seisdedos H, Ibarra-Molero B, Sanchez-Ruiz JM: **How many ionizable groups can sit on a protein hydrophobic core?** *Proteins* 2011, **80(1)**:1-7.
6. Xu J, Baase WA, Baldwin E, Matthews BW: **The response of T4 lysozyme to large-to-small substitutions within the core and its relation to the hydrophobic effect.** *Protein science : a publication of the Protein Society* 1998, **7(1)**:158-177.
7. Gong S, Blundell TL: **Structural and functional restraints on the occurrence of single amino acid variations in human proteins.** *PLoS ONE* 2010, **5(2)**:e9186.
8. Sunyaev S, Ramensky V, Bork P: **Towards a structural basis of human non-synonymous single nucleotide polymorphisms.** *Trends Genet* 2000, **16(5)**:198-200.
9. Wang Z, Moult J: **SNPs, protein structure, and disease.** *Human mutation* 2001, **17(4)**:263-270.
10. De Filippis V, Sander C, Vriend G: **Predicting local structural changes that result from point mutations.** *Protein engineering* 1994, **7(10)**:1203-1208.
11. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE: **The Protein Data Bank.** *Nucleic Acids Res* 2000, **28(1)**:235-242.
12. Li W, Godzik A: **Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences.** *Bioinformatics* 2006, **22(13)**:1658-1659.
13. Mika S, Rost B: **UniqueProt: Creating representative protein sequence sets.** *Nucleic Acids Res* 2003, **31(13)**:3789-3791.
14. Kabsch W, Sander C: **Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features.** *Biopolymers* 1983, **22(12)**:2577-2637.
15. McLachlan A: **Rapid comparison of protein structures.** *Acta Crystallographica Section A* 1982, **38(6)**:871-873.
16. **ProFit.** [http://www.bioinf.org.uk/software/profit/].
17. Bromberg Y, Rost B: **SNAP: predict effect of non-synonymous polymorphisms on function.** *Nucleic Acids Res* 2007, **35(11)**:3823-3835.
18. Kawabata T, Ota M, Nishikawa K: **The Protein Mutant Database.** *Nucleic Acids Res* 1999, **27(1)**:355-357.
19. 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 issue)**:W306-310.
20. Kumar MD, Bava KA, Gromiha MM, Prabakaran P, Kitajima K, Uedaira H, Sarai A: **ProTherm and ProNIT: thermodynamic databases for proteins and protein-nucleic acid interactions.** *Nucleic Acids Res* 2006, **34(Database issue)**:D204-206.
21. Capriotti E, Fariselli P, Rossi I, Casadio R: **A three-state prediction of single point mutations on protein stability changes.** *BMC bioinformatics* 2008, **9(Suppl 2)**:S6.
22. Fan R-E, Chang K-W, Hsieh C-J, Wang X-R, Lin C-J: **LIBLINEAR: A Library for Large Linear Classification.** *J Mach Learn Res* 2008, **9**:1871-1874.
23. Rost B, Yachdav G, Liu J: **The PredictProtein server.** *Nucleic Acids Res* 2004, **32(Web Server issue)**:W321-326.
24. Rost B: **PHD: Predicting one-dimensional protein structure by profile-based neural networks.** In *Methods in enzymology. Volume 266.* Academic Press;Russell FD 1996:525-539.
25. Rost B: **How to Use Protein 1- D Structure Predicted by PROFphd.** *The Proteomics Protocols Handbook* 2005, 875-901.
26. Zamyatnin AA: **Protein volume in solution.** *Progress in biophysics and molecular biology* 1972, **24**:107-123.
27. Kyte J, Doolittle RF: **A simple method for displaying the hydropathic character of a protein.** *Journal of molecular biology* 1982, **157(1)**:105-132.
28. Betts MJ, Russell RB: **Amino acid properties and consequences of substitutions.** *Bioinformatics for Geneticists* 2003, **317.**
29. 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.
30. Bairoch A, Apweiler R, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, *et al*: **The Universal Protein Resource (UniProt).** *Nucleic Acids Res* 2005, **33(Database issue)**:D154-159.
31. 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.
32. Rost B, Sander C: **Combining evolutionary information and neural networks to predict protein secondary structure.** *Proteins* 1994, **19(1)**:55-72.
33. Rost B, Sander C: **Prediction of protein secondary structure at better than 70% accuracy.** *Journal of molecular biology* 1993, **232(2)**:584-599.
34. Schlessinger A, Yachdav G, Rost B: **PROFbval: predict flexible and rigid residues in proteins.** *Bioinformatics* 2006, **22(7)**:891-893.
35. Schlessinger A, Liu J, Rost B: **Natively unstructured loops differ from other loops.** *PLoS computational biology* 2007, **3(7)**:e140.
36. Schlessinger A, Punta M, Rost B: **Natively unstructured regions in proteins identified from contact predictions.** *Bioinformatics* 2007, **23(18)**:2376-2384.
37. Dosztanyi Z, Csizmok V, Tompa P, Simon I: **IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content.** *Bioinformatics* 2005, **21(16)**:3433-3434.
38. Schlessinger A, Punta M, Yachdav G, Kajan L, Rost B: **Improved disorder prediction by combination of orthogonal approaches.** *PLoS ONE* 2009, **4(2)**:e4433.
39. Ofran Y, Rost B: **ISIS: interaction sites identified from sequence.** *Bioinformatics* 2007, **23(2)**:e13-16.
40. Ofran Y, Rost B: **Protein-protein interaction hotspots carved into sequences.** *PLoS computational biology* 2007, **3(7)**:e119.
41. Ofran Y, Rost B: **Analysing six types of protein-protein interfaces.** *Journal of molecular biology* 2003, **325(2)**:377-387.
42. Ofran Y, Mysore V, Rost B: **Prediction of DNA-binding residues from sequence.** *Bioinformatics* 2007, **23(13)**:i347-353.
43. Finn RD, Mistry J, Tate J, Coggill P, Heger A, Pollington JE, Gavin OL, Gunasekaran P, Ceric G, Forslund K, *et al*: **The Pfam protein families database.** *Nucleic Acids Res* 2010, **38(Database issue)**:D211-222.

44. Finn RD, Clements J, Eddy SR: **HMMER web server: interactive sequence similarity searching.** *Nucleic Acids Res* 2011, **39(Web Server issue)**:W29-37.
45. Smialowski P, Frishman D, Kramer S: **Pitfalls of supervised feature selection.** *Bioinformatics* 2010, **26(3)**:440-443.
46. Kabsch W, Sander C: **On the use of sequence homologies to predict protein structure: identical pentapeptides can have completely different conformations.** *Proceedings of the National Academy of Sciences of the United States of America* 1984, **81(4)**:1075-1078.
47. Cerpa R, Cohen FE, Kuntz ID: **Conformational switching in designed peptides: the helix/sheet transition.** *Folding & design* 1996, **1(2)**:91-101.
48. Fliess A, Motro B, Unger R: **Swaps in protein sequences.** *Proteins* 2002, **48(2)**:377-387.

doi:10.1186/1471-2164-13-S4-S4
Cite this article as: Schaefer and Rost: **Predict impact of single amino acid change upon protein structure.** *BMC Genomics* 2012 **13**(Suppl 4):S4.

Schaefer et al. BMC Genomics 2012, 13(Suppl 4):S11
http://www.biomedcentral.com/1471-2164/13/S4/S11

BMC Genomics

PROCEEDINGS Open Access

Disease-related mutations predicted to impact protein function

Christian Schaefer[1,2*], Yana Bromberg[5], Dominik Achten[1], Burkhard Rost[1,2,3,4]

From SNP-SIG 2011: Identification and annotation of SNPs in the context of structure, function and disease
Vienna, Austria. 15 July 2011

Abstract

Background: Non-synonymous single nucleotide polymorphisms (nsSNPs) alter the protein sequence and can cause disease. The impact has been described by reliable experiments for relatively few mutations. Here, we study predictions for functional impact of disease-annotated mutations from OMIM, PMD and Swiss-Prot and of variants not linked to disease.

Results: Most disease-causing mutations were predicted to impact protein function. More surprisingly, the raw predictions scores for disease-causing mutations were higher than the scores for the function-altering data set originally used for developing the prediction method (here SNAP). We might expect that diseases are caused by change-of-function mutations. However, it is surprising how well prediction methods developed for different purposes identify this link. Conversely, our predictions suggest that the set of nsSNPs not currently linked to diseases contains very few strong disease associations to be discovered.

Conclusions: Firstly, annotations of disease-causing nsSNPs are on average so reliable that they can be used as proxies for functional impact. Secondly, disease-causing nsSNPs can be identified very well by methods that predict the impact of mutations on protein function. This implies that the existing prediction methods provide a very good means of choosing a set of suspect SNPs relevant for disease.

Background

Evolution leads to genetic diversity

The selection of survival under changing conditions guides the cell's genetic makeup ("genotype") that is dynamically fit for retaining important cellular functions ("phenotype"). Today's genetic landscape represents the current state of a sampling process that continuously creates new phenotypes. This process yields genetic variation across and within species. In human, single nucleotide polymorphisms (SNPs) are essential for genetic diversity [1,2]. Non-synonymous SNPs (nsSNPs) alter the amino acid sequence. Some of these mutations affect protein structure and/or function and could increase susceptibility to disease.

* Correspondence: schaefer@rostlab.org
[1]TUM, Bioinformatics - i12, Informatics, Boltzmannstrasse 3, 85748 Garching/Munich, Germany
Full list of author information is available at the end of the article

Do disease-causing mutations impact protein function?

Disease-causing mutations occur often inside the protein (buried) and at hydrogen-bonding residues [3-5]. Protein function is often associated with evolutionarily conserved residues [4,6-9]. Most known disease-related nsSNPs in proteins of known 3D (three-dimensional) structure appear to affect structurally important residues and sites relevant for function [4]. For instance, disease-associated mutations can affect protein interactions [10]. In protein kinases, they have been shown to cluster into the functionally important catalytic core [11,12]. The above trends confirm the expectation that mutations cause disease because they damage important proteins.

Experts have established the above trends by laboriously inspecting small sets of well-curated proteins. Could less well-versed experts with better algorithms have established valid trends about disease-causing mutations for large data set by automatically extracting data set of disease-related mutations and their *predicted* functional effects? At

OMIM's infancy, a few years ago, we failed to accomplish this; i.e. observed trends did not differ much from random. This has changed. Here, we provide data that strongly suggest an affirmative answer to the question and demonstrate that we have a large repository of disease-causing mutations. To pick the most important practical result of our work: today's disease-causing mutations can serve as an excellent proxy for "change of function".

Methods

Data sets

We used SNPdbe [13] as the underlying source for amino acid substitutions, functional effect annotations and disease relations. This comprehensive new resource integrates variants from dbSNP [14], Swiss-Prot [15], PMD [16], and OMIM [17] and annotations of functional effects (from Swiss-Prot and PMD) and disease (from SwissVar [18], PMD and OMIM). The term 'genetic disease' is rather heterogeneous, covering Mendelian, monogenic disorders and polygenic diseases, exhibiting more complex genotypic patterns. Here, we do not differentiate between the different disease-types. Instead we aim at analyzing all disease-causing mutations.

We created the following five subsets from SNPdbe (Additional file 2). (1) *Set of disease-related + observed effect mutations:* We collected 1,105 human nsSNPs (from 217 proteins) that were annotated to be both disease-causing and functionally non-neutral. (2) *Set of disease-related mutations:* We obtained a set of amino acid substitutions in human proteins with disease-association. We extracted 26,404 mutations (3,419 proteins) with disease annotations but no annotated functional effect. (3) *Set of observed effect mutations:* We collected 36,317 mutants in 3,790 proteins with experimentally observed effect. We excluded mutations with disease associations. This set constitutes a part of the "functional effects" sets annotated in PMD; it served as the positive training set for SNAP [19]. Note that after our filtering the resulting set of mutations with *observed effect* and the set of *disease-related* mutants did NOT overlap. (*4*) *Set of mutations with unknown disease relation:* We extracted 251,414 variants (28,913 proteins) without known disease associations. (5) *Set of random mutations:* We randomly selected one mutation in each of the 28,913 proteins from the set of mutants of *unknown disease relation* such that the mutated position was maximally distant from any other mutation observed in the given protein.

Prediction of effect

For the vast majority of point mutants (single amino acid changes or nsSNPs) in human, the impact on protein function remains unknown. For all mutations in the above four data sets (disease-causing, disease-relation unknown, observed function-changing, and random), we predicted their effects on function with SNAP [19] and SIFT [20]. Both methods provide binary classifications (effect/neutral) along with a more detailed score. SNAP scores range from -100 (strongly predicted as neutral) to 100 (strongly predicted to change function); the distance from the binary decision boundary (0) measures the reliability of the effect. Essentially, stronger predictions are also more reliable, *i.e.* the higher the score, the more likely the mutation impacts function [19,21,22]. For a small data set, we previously established that SNAP scores correlate with the severity of change; *i.e.* high (positive) SNAP scores relate to more severe functional effects [19,21,22].

SIFT [20] scores range from 0 to 1 and aim at characterizing the normalized probability of tolerable amino acid substitution. Values $\leq$0.05 imply prediction of functional change; all other values are considered neutral. As with many other prediction methods, the distance to the decision boundary (0.05) reflects the reliability of a particular prediction [23]. For many prediction methods developed in our group (protein-protein binding [24-26], protein-DNA binding [27], backbone flexibility [28]), the strength of an effect correlated with prediction strength, *e.g.* ISIS predicted binding hot spots stronger than other residues involved in the interaction [26]. Although we never used the strength of an effect to train our methods, this correlation is intuitive: stronger effects are more consistent and therefore become stronger carved into the machine-learning model. Similarly, SIFT scores could be used to prioritize amino acid substitutions [23]. In this perspective, we consider the distance from the default decision boundary (0.05) as the magnitude of the effect.

SNAP and SIFT aspire to solve the same problem with different means. SNAP was trained on literature-derived [16] mutants that are either functionally similar to the wild-type (neutral) or alter function (effect) in either direction (*decrease* and *increase* of function). SIFT on the other hand infers probabilities of functional change from residue conservation in alignments of evolutionarily related proteins. While SNAP operates on an experimentally substantiated definition of change, SIFT uses conservation scores of amino acids as a proxy for functional change. Although both methods largely capture the underlying biological meaning of functional change, their predictions disagree often. Thus, the methods are likely orthogonal, picking up different aspects of protein function.

In addition, we applied PhD-SNP [29] to predict whether mutations in all five sets are disease-causing or neutral. PhD-SNP offers several modes striking different balances between runtime and performance. We used the most accurate mode that uses both sequence and evolutionary profiles.

Box plots

We represented our resulting distributions using box plots [30,31]. The lower and upper box edges depict the first and third quartiles of the distributions, respectively. The length of the box is the interquartile range of the distribution. The bold bar inside the box represents the median, while dashed lines reach to the most extreme data points, that are no more than 1.5 times the interquartile range away from the upper or lower box edge. Note that each box covers half the distribution.

Results and discussion

Disease-causing mutations strongly predicted to change protein function

We applied SNAP and SIFT to the 26,404 annotated *disease related* mutants (Methods). At the default threshold, SNAP predicted over 86% of the *disease related* mutations to impact function (Fig. 1A, B, 2) and SIFT ~59% (Fig. 2, Additional file 1). SNAP predictions were very strong: about half of the effect predictions had levels of severity of >40 (Fig. 1B, dashed black curve).

In our experience, SNAP scores >40 are exceptional when applying the method to new data. To clarify this point, the *observed effect* mutations were the very same data set that trained SNAP. We ascertained that this set had no overlap with the *disease related* mutations (Methods). Usually, machine-learning methods perform much better on the training than on the testing set. This also holds for SNAP; hence, the distribution of SNAP scores for the training set of *observed effect* mutants is expected to be closer to 'more effect' than for any other data set. We observed the opposite (Fig. 1B: red vs. dashed black lines): effect predictions were stronger for the *disease related* mutations than for our *observed effect* training set, *e.g.* while just over 40% of the training set reached a score >40, 47% of the *disease related* mutations did. A difference of seven percentage points might not be perceived as high, but the effect is significantly higher for comparison to testing on the training set. SIFT overall also predicted the *disease related* mutations stronger than the *observed effect* data, but the difference was not significant (Additional file 1).

Do disease-related mutations *with* an observed effect alter function even more? We analyzed the predicted functional effect of disease-associated mutations *with* observed effect (*disease-related+observed effect*). About 90% were predicted to impact function (4% more than for *disease related*), while over 53% had SNAP scores higher than 40 (6% more than for *disease related*; Fig. 1A, B solid black line, Fig. 2). SIFT showed a similar trend: 66% in the set of *disease related+observed effect* compared to 59% in *disease related* mutations (Fig. 2, Additional file 1). This suggests that the most reliable source of impact mutations is by connecting disease relations and independent experimental observations.

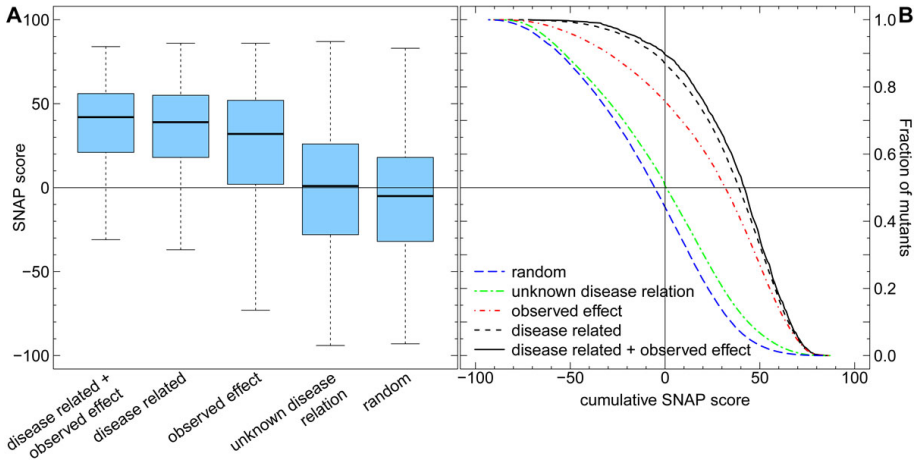

Figure 1 Disease-causing mutations have highest scores SNAP predicted the impact of function for five different data sets of point mutations: *disease related + observed effect* and *disease related* mutants, mutations with *observed effect*, *unknown disease relation*, and *random* mutations. For each set we display the predicted functional severity of mutations. (A) Scores above zero (horizontal line) correspond to *effect*, scores below to *neutral*, the distance from 0 correlates to severity; lower/upper bound and bar in the box represent the lower/upper quartile and median. 90% of *disease related+observed effect* and over 86% of the *disease related* mutations were predicted to effect function, compared to only 51% in mutations of *unknown disease relation*. Effect predictions dominated the *observed effect* mutants less (76%) than the *disease related* mutants (86%). The effect in *random* mutations (44%) provided an upper bound for effect mutations in proven non-disease related variants. (B) Cumulative distributions of predicted functional severity; points on a curve correspond to fractions (y-axis) of mutations with SNAP scores (x-axis) ≥ this value. The vertical line separates *neutral* from *effect*. Disease-causing mutations were predicted to be most severe (black solid and dashed lines above all others). These results suggest that change in function may explain most disease-related mutations.

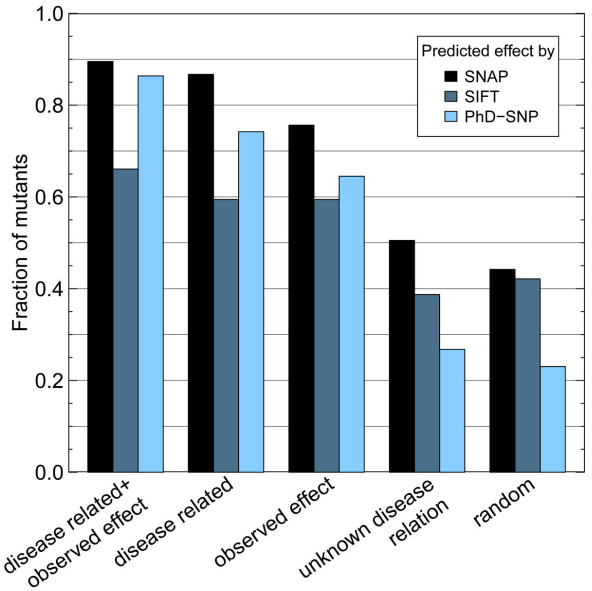

Figure 2 Effect most prevalent in disease mutants For each set we show the fraction of mutants with predicted effect (SNAP, SIFT: functional effect, PhD-SNP: disease). Disease predictions taken from PhD-SNP (light blue bars) confirm the major observation found in functional predictions (black+dark blue bars): *observed effect* mutants have high impact on disease. More than 64% of these are predicted to be disease-causing while only 27% of mutations of *unknown disease relation* are predicted to cause disease.

As negative control, the predictions differed greatly for the 251,414 mutants with *unknown disease relation*. First, only about 51% of those were predicted to have an effect by SNAP (Fig. 1A, B, 2), and only 39% by SIFT (Fig. 2, Additional file 1). Second, only 12% of those had a SNAP score larger than 40 (Fig. 1B, dashed green curve).

Many mutations with unknown effect predicted to alter function

SNAP and SIFT predicted much more effect for *disease related* mutations than in mutants with *unknown disease relation*. Still, many of those mutations were predicted to change protein function. However, much fewer mutants with *unknown disease relation* were predicted to significantly change function than the *disease related* mutations (Fig. 1B: strong effect for 14% of mutants *unknown disease relation* - dashed green line - vs. 48% of *disease related* mutations - dashed black line). Comparing the prediction trends between the two data sets suggests that the mutations of *unknown disease relation* will never become a 'disease-rich' set (i.e. through newly discovered disease associations). R*andom* mutations were even less often predicted to have strong effect (~7%, Fig. 1B, dashed blue line). This result suggests that many experimental annotations of 'functional impact' remain to be determined/observed for the set of mutations with *unknown disease relation* (roughly > 7%-14%).

Same trend found in predicted disease mutations

If *disease related* can serve as a good proxy for (strong) functional impact, then a method trained to predict disease-causing mutations should reveal the reverse and thus confirm the same: predicted disease is expected to be enriched in *observed effect* compared to mutations of *unknown disease relation*. We analyzed the fraction of predicted disease by applying PhD-SNP (Methods) to our five data sets. PhD-SNP predicted >64% of the o*bserved effect* mutations as disease related (Fig. 2), while only 26% of mutations with *unknown disease relation* were predicted to be disease associated. Furthermore, we confirmed the other observations already found in functional impact predictions: Random mutations appear to have the lowest impact on disease (only 22%, Fig. 2).

PhD-SNP predicted both disease-related sets to contain most disease mutants (86% in *disease related +observed effect* and 74% in *disease related*, Fig. 2). This was expected due to the important overlap between our data and the training set of PhD-SNP [29]. Nonetheless, the increase in predicted-disease mutations of 12% once again suggested that *observed effect* mutants play a major role in disease.

Our findings show that if a mutation leads to disease then a change in function plays a major role in explaining the cause (59%-86%). This finding cannot be inverted due to the overlap of score distributions of *disease related* mutants and mutants with *unknown disease annotation* (Fig. 1A, Additional file 1); i.e. strong effect on function does not imply disease.

Our comparison between mutations annotated as *disease related* and those experimentally annotated function changing (*observed effect*) does not imply that there is anything special about disease-causing mutations. Instead, our findings highlight differences in the *severity* of functional effect. That is, on average, assuming that a disease causing mutation has a functional effect is more reliable than experimentally evaluating functional change.

Conclusions

We compared disease-associated single point mutations (nsSNPs) predicted to change protein function with those of unknown disease-association. Implicitly, we tested the reliability of annotations that link mutations to disease and the extent to which predictions of functional effect overlap with disease causation.

As opposed to other studies addressing this question [3-6,10-12], we used predictions of functional effect to determine the fraction of deleterious point mutations in

two different populations of human variants: *disease related* (or disease-causing) mutations and mutations without any knowledge of phenotypic effect. The major findings were: (1) annotations of disease-causation provide a good approximation of functional effect. (2) Methods developed to predict the impact of mutations onto protein function clearly identify disease-causing mutations as those that change function. In other words, their predictions provide a valuable first step towards the study of the molecular impact of disease.

Funding

This work 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); YB was supported by the SEBS, Rutgers, New Brunswick startup funds.

Additional material

Additional file 1: SIFT predictions. Non-neutral mutations are enriched in a set of disease-causing variants, whereas they are depleted in variants with no known linkage to disease.

Additional file 2: Mutation and sequence data. Archive of the five different mutant sets used in this study separated by SNAP/SIFT and PhD-SNP predictions including the protein wild type sequences.

Acknowledgements

Special thanks to Laszlo Kajan (TUM), Guy Yachdav (TUM/Columbia University), and Tim Karl (TUM) for help with software and hardware; to Marlena Drabik (TUM) for administrative support. Thanks to Rolf Apweiler (UniProt, EBI, Hinxton), Amos Bairoch (CALIPHO, SIB, Geneva), Ioannis Xenarios (Swiss-Prot, SIB, Geneva), their crews, and those from OMIM, PMD, SwissVar and dbSNP for maintaining excellent databases. Last, but not least, thanks to all experimentalists who enabled this analysis by making their data publicly available.

This article has been published as part of *BMC Genomics* Volume 13 Supplement 4, 2012: SNP-SIG 2011: Identification and annotation of SNPs in the context of structure, function and disease. The full contents of the supplement are available online at http://www.biomedcentral.com/bmcgenomics/supplements/13/S4.

Author details

[1]TUM, Bioinformatics - i12, Informatics, Boltzmannstrasse 3, 85748 Garching/Munich, Germany. [2]TUM Graduate School of Information Science in Health (GSISH), Boltzmannstr. 11, 85748 Garching/Munich, Germany. [3]Institute of Advanced Study (TUM-IAS), Lichtenbergstr. 2a, 85748 Garching/Munich, Germany. [4]Columbia University, Department of Biochemistry and Molecular Biophysics & New York Consortium on Membrane Protein Structure (NYCOMPS), 701 West, 168th Street, New York, NY 10032, USA. [5]Department of Biochemistry and Microbiology, School of Environmental and Biological Sciences, Rutgers University, New Brunswick, NJ 08901, USA.

Authors' contributions

CS participated in the design of the study, performed the data analysis and helped to draft the manuscript. YB participated in the design of the study and helped draft the manuscript. DA participated in the design of the study. BR participated in the coordination and design of the study and helped to draft the manuscript.

Competing interests

The authors declare they have no competing interests.

Published: 18 June 2012

References

1. Consortium GP: **A map of human genome variation from population-scale sequencing.** *Nature* 2010, **467(7319)**:1061-1073.
2. Collins FS, Brooks LD, Chakravarti A: **A DNA polymorphism discovery resource for research on human genetic variation.** *Genome Res* 1998, **8(12)**:1229-1231.
3. Gong S, Blundell TL: **Structural and functional restraints on the occurrence of single amino acid variations in human proteins.** *PLoS ONE* 2010, **5(2)**:e9186.
4. Sunyaev S, Ramensky V, Bork P: **Towards a structural basis of human non-synonymous single nucleotide polymorphisms.** *Trends Genet* 2000, **16(5)**:198-200.
5. Wang Z, Moult J: **SNPs, protein structure, and disease.** *Hum Mutat* 2001, **17(4)**:263-270.
6. Talavera D, Taylor MS, Thornton JM: **The (non)malignancy of cancerous amino acidic substitutions.** *Proteins* 2010, **78(3)**:518-529.
7. Lichtarge O, Bourne HR, Cohen FE: **An evolutionary trace method defines binding surfaces common to protein families.** *J Mol Biol* 1996, **257(2)**:342-358.
8. Rausell A, Juan D, Pazos F, Valencia A: **Protein interactions and ligand binding: from protein subfamilies to functional specificity.** *Proc Natl Acad Sci U S A* 2010, **107(5)**:1995-2000.
9. Landgraf R, Xenarios I, Eisenberg D: **Three-dimensional cluster analysis identifies interfaces and functional residue clusters in proteins.** *J Mol Biol* 2001, **307(5)**:1487-1502.
10. Schuster-Bockler B, Bateman A: **Protein interactions in human genetic diseases.** *Genome Biol* 2008, **9(1)**:R9.
11. Torkamani A, Schork NJ: **Distribution analysis of nonsynonymous polymorphisms within the human kinase gene family.** *Genomics* 2007, **90(1)**:49-58.
12. Torkamani A, Verkhivker G, Schork NJ: **Cancer driver mutations in protein kinase genes.** *Cancer Lett* 2009, **281(2)**:117-127.
13. Schaefer C, Meier A, Rost B, Bromberg Y: **SNPdbe: constructing an nsSNP functional impacts database.** *Bioinformatics* 2012, **28(4)**:601-602.
14. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K: **dbSNP: the NCBI database of genetic variation.** *Nucleic Acids Res* 2001, **29(1)**:308-311.
15. Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O'Donovan C, Phan I, *et al*: **The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003.** *Nucleic Acids Res* 2003, **31(1)**:365-370.
16. Kawabata T, Ota M, Nishikawa K: **The Protein Mutant Database.** *Nucleic Acids Res* 1999, **27(1)**:355-357.
17. Amberger J, Bocchini CA, Scott AF, Hamosh A: **McKusick's Online Mendelian Inheritance in Man (OMIM).** *Nucleic Acids Res* 2009, **37(Database issue)**:D793-796.
18. Mottaz A, David FP, Veuthey AL, Yip YL: **Easy retrieval of single amino-acid polymorphisms and phenotype information using SwissVar.** *Bioinformatics* 2010, **26(6)**:851-852.
19. Bromberg Y, Rost B: **SNAP: predict effect of non-synonymous polymorphisms on function.** *Nucleic Acids Res* 2007, **35(11)**:3823-3835.
20. Ng PC, Henikoff S: **Predicting deleterious amino acid substitutions.** *Genome Res* 2001, **11(5)**:863-874.
21. 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(9)**:3059-3069.
22. Bromberg Y, Rost B: **Comprehensive in silico mutagenesis highlights functionally important residues in proteins.** *Bioinformatics* 2008, **24(16)**: i207-212.
23. Ng PC, Henikoff S: **SIFT: Predicting amino acid changes that affect protein function.** *Nucleic Acids Res* 2003, **31(13)**:3812-3814.
24. Ofran Y, Rost B: **Analysing six types of protein-protein interfaces.** *J Mol Biol* 2003, **325**:377-387.
25. Ofran Y, Rost B: **ISIS: interaction sites identified from sequence.** *Bioinformatics* 2007, **23(2)**:e13-16.
26. Ofran Y, Rost B: **Protein-protein interaction hot spots carved into sequences.** *PLoS Computational Biology* 2007, **3(7)**:e119.
27. Ofran Y, Mysore V, Rost B: **Prediction of DNA-binding residues from sequence.** *Bioinformatics* 2007, **23(13)**:i347-353.

28. Schlessinger A, Yachdav G, Rost B: **PROFbval: predict flexible and rigid residues in proteins.** *Bioinformatics* 2006, **22(7)**:891-893.
29. 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.
30. McGill R, Tukey JW, Larsen WA: **Variations of Box Plots.** *The American Statistician* 1978, **32(1)**:12-16.
31. Tukey JW: **Exploratory data analysis.** Reading, Mass.: Addison-Wesley Pub. Co.; 1977.

doi:10.1186/1471-2164-13-S4-S11
Cite this article as: Schaefer *et al.*: Disease-related mutations predicted to impact protein function. *BMC Genomics* 2012 **13**(Suppl 4):S11.