

# Pathogenic or Not? And If So, Then How? Studying the Effects of Missense Mutations Using Bioinformatics Methods

Janita Thusberg<sup>1</sup> and Mauno Vihinen<sup>1,2\*</sup>

<sup>1</sup>*Institute of Medical Technology, FI-33014 University of Tampere, Finland;* <sup>2</sup>*Tampere University Hospital, FI-33520 Tampere, Finland*

Communicated by Mark H. Paalman

Received 26 July 2008; accepted revised manuscript 9 October 2008.

Published online 6 March 2009 in Wiley InterScience (www.interscience.wiley.com). DOI 10.1002/humu.20938

**ABSTRACT:** Many gene defects are relatively easy to identify experimentally, but obtaining information about the effects of sequence variations and elucidation of the detailed molecular mechanisms of genetic diseases will be among the next major efforts in mutation research. Amino acid substitutions may have diverse effects on protein structure and function; thus, a detailed analysis of the mutations is essential. Experimental study of the molecular effects of mutations is laborious, whereas useful and reliable information about the effects of amino acid substitutions can readily be obtained by theoretical methods. Experimentally defined structures and molecular modeling can be used as a basis for interpretation of the mutations. The effects of missense mutations can be analyzed even when the 3D structure of the protein has not been determined, although structure-based analyses are more reliable. Structural analyses include studies of the contacts between residues, their implication for the stability of the protein, and the effects of the introduced residues. Investigations of steric and stereochemical consequences of substitutions provide insights on the molecular fit of the introduced residue. Mutations that change the electrostatic surface potential of a protein have wide-ranging effects. Analyses of the effects of mutations on interactions with ligands and partners have been performed for elucidation of functional mutations. We have employed numerous methods for predicting the effects of amino acid substitutions. We discuss the applicability of these methods in the analysis of genes, proteins, and diseases to reveal protein structure–function relationships, which is essential to gain insights into disease genotype–phenotype correlations.

Hum Mutat 30:703–714, 2009. © 2009 Wiley-Liss, Inc.

**KEY WORDS:** missense mutation; mutation analysis; bioinformatics; computational methods; effects of mutations; structural basis of disease

## Introduction

The knowledge of the complete human genome sequence and the rapid accumulation of variation data allow a more mechanism-based approach to the understanding of the relationship between genotype and disease. With powerful strategies for elucidating genetic defects such as whole genome association studies and high-throughput, low-cost sequencing, genotyping ceases to be the bottleneck for the understanding of genetic disease. Gene defects are being identified at an increasing pace, and obtaining information about the effects of sequence variation and elucidation of the detailed molecular mechanisms of genetic disease will be the next major efforts in mutation research. The effects of large changes, such as gross deletions or insertions, are relatively easy to explain, but the consequences of missense mutations require more detailed study at the protein level.

There are about 10 million single nucleotide polymorphisms (SNPs) in the human genome that have an appreciable frequency (i.e., >1%) [The International HapMap Consortium, 2003], of which 67,000–200,000 have been estimated to be nonsynonymous coding SNPs (nsSNPs) [Cargill et al., 1999; Halushka et al., 1999; Livingston et al., 2004]. A nonsynonymous, missense variant is a single base change in a coding region that causes an amino acid change in the corresponding protein. Missense mutations, in contrast to SNPs, are rather rare events. However, numerous single gene diseases have been attributed to missense mutations. Testing of the possible association of all the nonsynonymous genetic variants with disease or experimental characterization of their effects on protein function would be extremely expensive, time consuming, and difficult—especially in diseases that are caused by a large and varying number of mutations, such as cancer. The computational study of their putative effects would be beneficial in prioritizing the most probable disease-causing variations for association with diseases. On the other hand, those missense mutations already known to be associated with disease can be studied computationally in order to identify pharmaceutical targets for relevant treatments and to gain insight into the molecular disease mechanisms. Predicting the effects of amino acid substitutions is also essential for the rational design of novel proteins by site-directed mutagenesis.

A disease phenotype may arise when an amino acid substitution affects a residue critical in protein function, for example, a residue in the catalytic site of an enzyme or a residue involved in crucial interactions with partner molecules. Alongside with the diseases caused by mutations leading to loss of function, gain of function may result from irregular or tighter binding of ligands or loss of

\*Correspondence to: Mauno Vihinen, University of Tampere, Institute of Medical Technology, Tampere, FI-33014, Finland. E-mail: mauno.vihinen@uta.fi

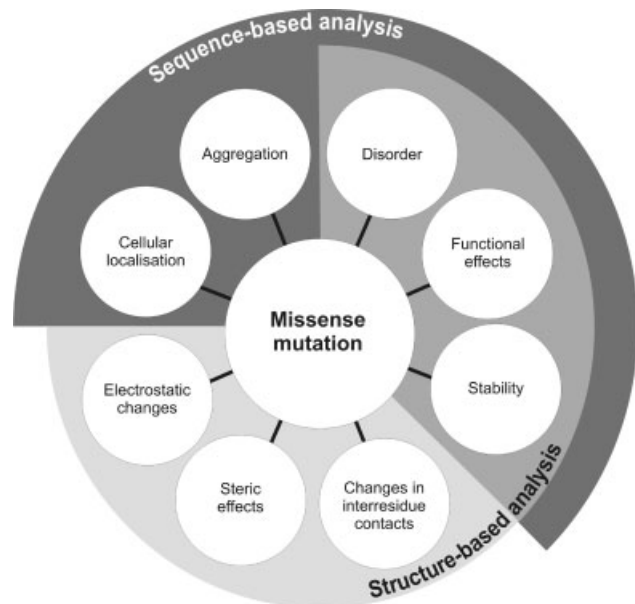
Contract grant sponsors: The Tampere Graduate School in Biomedicine and Biotechnology; Academy of Finland; The Medical Research Fund of the Tampere University Hospital.

specificity of a protein. In addition to the direct functional effects a substitution may have, a missense mutation may also lead to alterations in the protein structural properties, causing abnormal folding, structural instability, or aggregation of the protein. Even minor changes in the size or chemical nature of an amino acid side chain can alter or prevent the function of the protein. On the other hand, protein molecules are rather robust, and allow insertions to numerous sites without any effect on protein function [Pajunen et al., 2007; Poussu et al., 2004]. Furthermore, missense mutations may affect protein posttranslational modifications, for example, by inserting or deleting phosphorylation or glycosylation sites or protease cleavage sites, or altering signals guiding cellular localization. It should be noted that in addition to the direct effects on the protein molecules discussed in this paper, genetic variations may also cause disease phenotypes by affecting pretranslational processes, such as altering transcriptional regulation, mRNA stability, mRNA splicing, or translation rates. According to the data for monogenic diseases in HGMD, all the pretranslational effects account for less than 10% of cases [Stenson et al., 2003]. However, although alterations in the structure, function, or expression of the protein often cause a disease, this is not always the case, given the multiple redundancies of cellular pathways.

Bioinformatics methods can be helpful at several steps of the analysis (Fig. 1). Mutation databases serve as a starting point, providing the data for the analysis. Databases often contain curated information about the phenotypic effects of the mutations, together with information about the gene and protein in question. Sequence analysis provides information about the sites that are conserved in evolution that often have a crucial role in protein structure or function. There are numerous sequence-based predictors available for the prediction of the effect of a mutation on various biochemical properties of a protein, such as aggregation propensity, disorder, or stability. When there is an experimentally determined structure available for the protein of interest, the mutation analysis can be taken to the structural level,

making the analysis more reliable and complete. Alternatively, a modeled structure can be used. The mutations can be modeled into the structure, and after optimizing the side chain angles the role of the new residue can be studied in the context of its surroundings. It can be seen whether the new side chain fits into the structure at all, and the effects of the amino acid substitution on side-chain interactions can be studied in detail. Many programs predicting the effects of mutations also require the 3D coordinates of the wild-type protein as input. Bioinformatics methods, despite being useful in providing information about the nature of mutations as such, may also be helpful in guiding the design of further experimental research.

Several recent studies have applied computational methods to predict potentially deleterious effects of nonsynonymous SNPs in humans [Chasman and Adams, 2001; Hyytinen et al., 2002; Lau and Chasman, 2004; Miller and Kumar, 2001; Ng and Henikoff, 2001; Sunyaev et al., 2001a, b; Terp et al., 2002; Torkamani and Schork, 2007; Wang and Moulton, 2001; Wood et al., 2007; Worth et al., 2007]. Until now, the research has mainly concentrated on using just one or a few methods in one study, but the emerging trend in mutation analysis is to utilize a more extensive set of prediction methods in order to attain more reliable results [Burke et al., 2007; Lappalainen et al., 2008; Tavtigian et al., 2008a, b; Thusberg and Vihinen, 2006, 2007; Worth et al., 2007]. In this paper we present the current methodology and services available for mutation analysis and discuss their applicability in the analysis of genes, proteins, and diseases to reveal protein structure–function relationships, which is essential to gain insights into disease genotype–phenotype correlations. The missense mutation analysis approach is based on our experience during the last 15 years in studying and interpreting mutations and their effects in numerous diseases, especially including immunodeficiencies and cancers [Lappalainen et al., 2000, 2008; Lappalainen and Vihinen, 2002; Rong et al., 2000; Rong and Vihinen, 2000; Thusberg and Vihinen, 2006, 2007; Vihinen et al., 1994a, b, 1995, 1999].



**Figure 1.** A schematic figure of the groups of methods for analyzing the effects of missense mutations. Our approach can be divided into sequence- and structure-based sections (dark gray and light gray backgrounds, respectively), which in part overlap.

## Methods for the Analysis of Mutations

### Databases

Mutation databases serve as the basis for bioinformatics research on the effects of mutations and the structural basis of diseases. Central mutation databases (CMDDBs), the most prominent being the Human Gene Mutation Database (HGMD) [Stenson et al., 2008] and Online Mendelian Inheritance in Man (OMIM) [Hamosh et al., 2005], collect variants in all genes, mainly from the literature. The UniProtKB/Swissprot database contains manually annotated protein entries that feature partial lists for known sequence variants [Yip et al., 2008]. There are also databases available that aim at annotating human variation data with phenotype variations and protein structural and functional information, such as MS2PH-db (<http://ms2phdb-pbil.ibcp.fr/cgi-bin/home>), MutDB [Dantzer et al., 2005], SAAPdb [Cavallo and Martin, 2005], and KMDB/MutationView [Minoshima et al., 2001]. Locus-specific databases (LSDBs) list variants in specific genes and are typically manually annotated. General recommendations for the generation and curation of such databases have been proposed [Cotton et al., 2008], and rules for nomenclature of mutations are discussed in [den Dunnen and Antonarakis, 2000]. The Human Genome Variation Society maintains a list of available LSDBs (around 700) and CMDDBs (19) on their Website (<http://www.hgvs.org/dblist/dblist.html>). Genome browsers, such as the

University of California, Santa Cruz (UCSC) Genome Browser [Kent et al., 2002], the National Center for Biotechnology Information (NCBI) Map Viewer [Wheeler et al., 2003], and the Ensembl Genome Browser [Stalker et al., 2004], can also be used to obtain information about genes, their products, and sequence variants. PhenCode [Giardine et al., 2007] is a service that connects human phenotype and clinical data in LSDBs with data from the UCSC Genome Browser.

## Sequence Conservation

Disease-causing mutations have been shown to be overabundant at evolutionarily conserved positions, because these positions are usually essential for the structure or function of the protein [Miller and Kumar, 2001; Mooney and Klein, 2002; Ng and Henikoff, 2003; Shen and Vihinen, 2004; Sunyaev et al., 2001b; Vitkup et al., 2003] (example in Fig. 2C), whereas there is a general underabundance of disease-associated mutations in positions that show any potential to change in evolution [Briscoe et al., 2004; Miller and Kumar, 2001]. Furthermore, the amino acid changes caused by disease-causing mutations are more radical in terms of the differences in their physicochemical properties from the wild-type amino acids, compared to the differences observed between species [Briscoe et al., 2004; Miller and Kumar, 2001; Tang et al., 2004]. For studying the pathogenicity of a missense mutation, knowledge of the level and type of evolutionary conservation of the position is valuable in order to gain insight into the possible role of that position in the structure or function of the protein (Fig. 2C), and what types of amino acids can be exchanged freely without negatively impacting protein function [Miller and Kumar, 2001] (Fig. 2B). In addition to the conservation of a particular amino acid in a sequence position, the physicochemical properties of the amino acids (e.g., hydrophathy, charge, size) can be conserved for structural integrity or function (Fig. 2B). Another mechanism of conservation is covariation, where a compensating mutation occurs at another position in the protein. Networks of covariant amino acids may reveal positions important for protein structure or function when the role of these positions is not obvious when looking at the protein structure, because the positions may be linked either functionally, energetically or by forming a physical interaction in some important conformation of the protein [Gloor et al., 2005; Lockless and Ranganathan, 1999; Suel et al., 2003]. The coupling of two sites in a protein should cause these two positions to coevolve [Lichtarge et al., 1996; Marcotte et al., 1999; Pellegrini et al., 1999].

There are numerous methods available for multiple sequence alignment (MSA) and subsequent analysis of sequence conservation. Classic methods such as ClustalW [Thompson et al., 1994] can give reasonably accurate results for similar sequences, but fail to produce accurate alignments for divergent sequences [Thompson et al., 1999]. Many efforts have been made to characterize the accuracy of the various MSA methods [Ahola et al., 2006; Golubchik et al., 2007; Nuin et al., 2006; Raghava et al., 2003], but the overall outcome of these studies is that a perfect MSA method does not exist and that individual methods have their specific strengths and weaknesses. This makes the choice of the most suitable alignment method difficult. There are services available for running several MSA methods and combining their output into a single model, for example, the M-Coffee Web server [Moretti et al., 2007]. The most widely used and state-of-the-art sequence alignment methods are listed in Table 1. Alternatively, a ready-made

sequence alignment can be obtained from the Pfam database [Finn et al., 2008].

There are several alternative methods for the detection of positional sequence conservation and identification of individual conserved residues within a position [Ahola et al., 2004]. The visualization of MSAs makes it convenient to interpret the information contained in them, for example, the visualization tools (see Table 1) calculate conservation indices for each position in the alignment, and add color codes into the alignment for different levels of sequence conservation. Some methods, for example, ConSurf [Glaser et al., 2003; Landau et al., 2005], apply the color-coding scheme to protein structures, so that the user can visualize the structure color coded by the level of conservation of individual residues. Physicochemical conservation of amino acids can be detected by those visualization methods that assign distinct colors for groups of each type of amino acid (e.g., hydrophobic, hydrophilic, charged) and display them according to their prevalence in the alignment. An example of this kind of tool is MultiDisp (P. Riikonen and M. Vihinen, in preparation) (Fig. 2B).

Calculation of mutual information between pairs of sites in the multiple sequence alignment and subsequent building of covariant networks of amino acids can be done by the methods aaMI [Gloor et al., 2005] or ProCon [Shen and Vihinen, 2004]. MatrixPlot [Gorodkin et al., 1999] is a method for generating mutual information plots for sequence alignments.

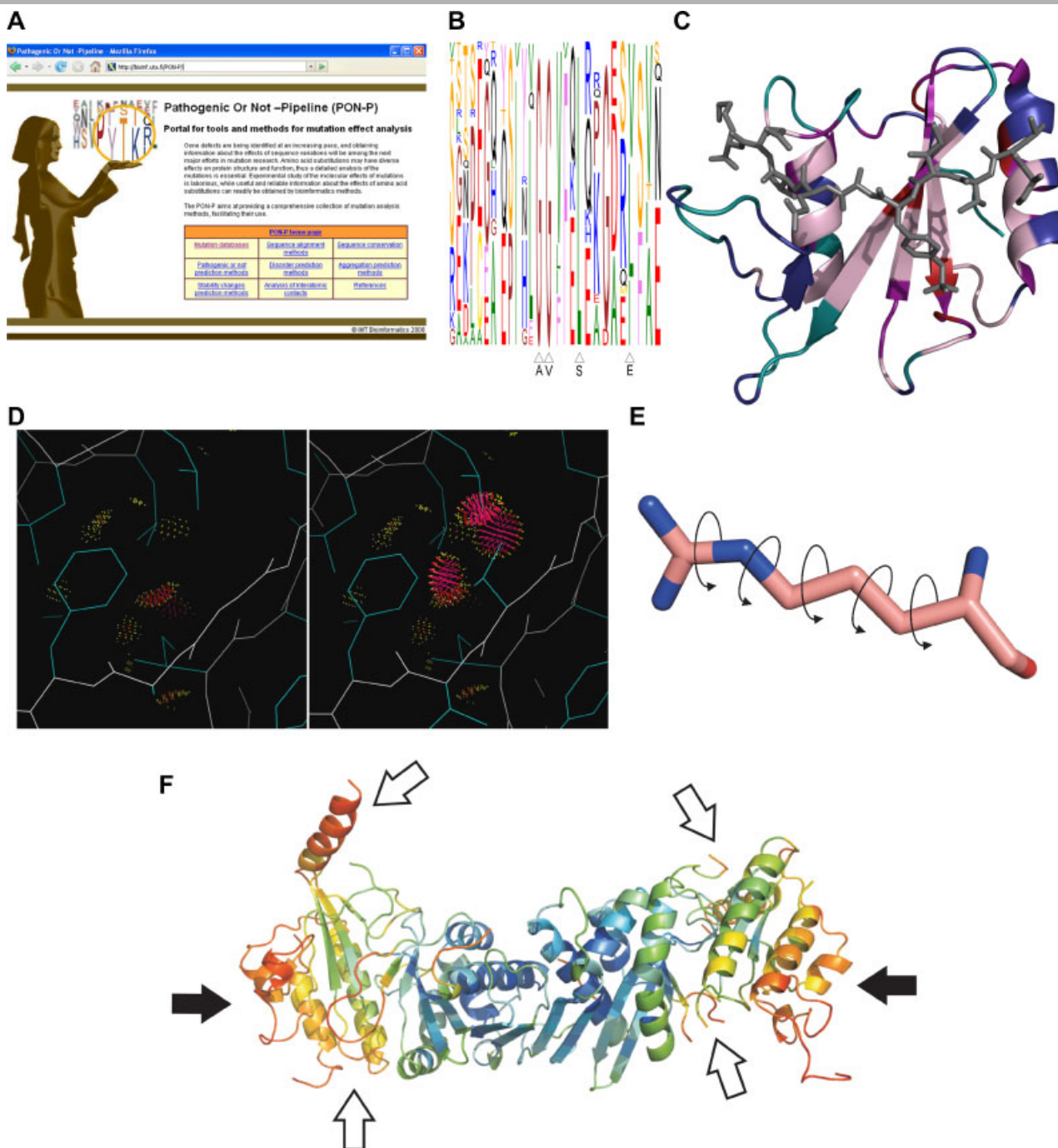
## Protein Localization

To function in its proper context, a protein must be translocated to the appropriate cellular compartment after translation. Proteins are typically directed to the right location by short peptide sequences that act as targeting signals. A missense mutation in the signal peptide might lead to the disruption or alteration of the signal. If the protein fails to be transported to the correct subcellular location, central reactions may be inactivated or signaling cascades misregulated. On the other hand, the mislocalized protein may be active in the wrong cellular compartment, causing harmful effects. Alterations to localization signals are rare, but the effects of mutations on them should be studied as part of the analysis of the effects of missense mutations [Laurila and Vihinen, submitted].

Several methods have also been developed for the prediction of the protein subcellular localization. These methods are discussed in detail in the review article by Schneider and Fechner [2004]. Recently, a protocol was introduced to combine several predictors [Emanuelsson et al., 2007], which was implemented by Laurila and Vihinen [submitted].

## Disorder

Many globular proteins contain segments that lack an ordered secondary structure, and some proteins even have global disorder, that is, do not fold in an ordered way. Instead of folding into fixed 3D structures, disordered proteins or protein segments exist as ensembles of interchanging structures (example in Fig. 2F). Intrinsically disordered proteins function in molecular recognition, molecular assembly/disassembly, protein modification, and entropic chains [Dunker et al., 2002], and they also have scavenger [Tomba, 2002] and chaperone [Tomba and Csermely, 2004] functions. Mutations may introduce disorder into usually ordered parts of a protein, thereby causing alterations in the protein fold, leading to possible changes in protein function. Increased flexibility of the protein may lead to differences in specificity, or



**Figure 2.** **A:** Screenshot of the Pathogenic-or-Not Pipeline (PON-P). **B:** A part of a MultiDisp visualization of the sequence alignment for CD40L and its homologs. The height of the characters indicates the frequency of the amino acids in the alignment positions, and the color of the objects reflects the chemical nature of the amino acids. Arrowheads below the alignment indicate positions of missense mutations in CD40L, together with mutant forms. Mutations are found in invariant positions and a charged residue (glutamic acid) is introduced in a position where hydrophobicity is the conserved amino acid property. **C:** The SH2D1A protein in complex with a phosphopeptide ligand (PDB ID 1D4W). The level of sequence conservation can give clues on the function of the protein. In the SH2 domains, the most conserved regions are involved in ligand binding. The color coding refers to sequence conservation in SH2 domains [Lappalainen et al., 2008]. The most conserved positions are colored red, followed by light pink, magenta, cyan, and the most variable regions are colored blue. The phosphopeptide ligand is colored gray. The figure is created by PyMOL [DeLano, 2002]. **D:** The substitution of G227 by V in CD40L causes serious clashes with the neighboring side chains. Left: wild-type protein. Right: mutated protein. Yellow—negligible overlap; red—significant overlap  $\geq 0.25$  Å; hot pink—serious clash overlap  $\geq 0.4$  Å. The figure is created by KiNG [Lovell et al., 2003]. **E:** Schematic representation of amino acid side chain  $\chi$  angle rotation. The arrows indicate the bonds that can be rotated over the full range of angles by the Bondrot function in Probe [Word et al., 1999, 2000]. **F:** Homodimeric structure of type II $\beta$  phosphatidylinositol phosphate kinase (PDB ID 1B01) coloured according to the B-values of individual residues (red—highest B-values, followed by orange, yellow, green, light blue, and dark blue—lowest B-values). The disordered regions in the protein are seen as missing electron densities (indicated by white arrows), surrounded by regions with high B-values. The C-terminal domains in each monomer have high thermal factors as well, because they are more flexible than the rest of the enzyme (black arrows) [Rao et al., 1998].

**Table 1. Methods for the Analysis of Missense Mutations and Their Effects**

Service name	URL	Description	Reference
Pathogenic or not predictors			
nsSNPAnalyzer	<a href="http://snpanalyzer.utmemb.edu/">http://snpanalyzer.utmemb.edu/</a>	Pathogenic or not	(Bao et al., 2005)
Panther	<a href="http://www.pantherdb.org/tools/csnpscoreForm.jsp">http://www.pantherdb.org/tools/csnpscoreForm.jsp</a>	Conservation analysis, pathogenic or not	(Thomas et al., 2003)
PhD-SNP	<a href="http://gpcr.biocomp.unibo.it/cgi/predictors/PhD-SNP/PhD-SNP.cgi">http://gpcr.biocomp.unibo.it/cgi/predictors/PhD-SNP/PhD-SNP.cgi</a>	Pathogenic or not	(Capriotti et al., 2006)
PMut	<a href="http://mmb2.pcb.ub.es:8080/PMut/">http://mmb2.pcb.ub.es:8080/PMut/</a>	Pathogenic or not	(Ferrer-Costa et al., 2005)
PolyPhen	<a href="http://coot.embl.de/PolyPhen/">http://coot.embl.de/PolyPhen/</a>	Pathogenic or not	(Ramensky et al., 2002)
SIFT	<a href="http://blocks.fhcr.org/sift/SIFT.html">http://blocks.fhcr.org/sift/SIFT.html</a>	Pathogenic or not	(Ng and Henikoff, 2001)
SNAP	<a href="http://cubic.bioc.columbia.edu/services/SNAP/">http://cubic.bioc.columbia.edu/services/SNAP/</a>	Pathogenic or not	(Bromberg and Rost, 2007)
SNPs3D	<a href="http://www.snps3d.org/">http://www.snps3d.org/</a>	Pathogenic or not	(Yue et al., 2006)
Sequence alignment methods			
M-Coffee	<a href="http://www.tcoffee.org/">http://www.tcoffee.org/</a>	Multiple sequence alignment	(Wallace et al., 2006)
MAFFT	<a href="http://align.bmr.kyushu-u.ac.jp/mafft/online/server/">http://align.bmr.kyushu-u.ac.jp/mafft/online/server/</a>	Multiple sequence alignment	(Katoh et al., 2002, 2005)
PROBCONS	<a href="http://probcons.stanford.edu/">http://probcons.stanford.edu/</a>	Multiple sequence alignment	(Do et al., 2005)
PROMALS	<a href="http://prodata.swmed.edu/promals/">http://prodata.swmed.edu/promals/</a>	Multiple sequence alignment	(Pei et al., 2007)
ClustalW2	<a href="http://www.ebi.ac.uk/Tools/clustalw2/index.html">http://www.ebi.ac.uk/Tools/clustalw2/index.html</a>	Multiple sequence alignment	(Larkin et al., 2007)
MUSCLE	<a href="http://www.ebi.ac.uk/Tools/muscle/index.html">http://www.ebi.ac.uk/Tools/muscle/index.html</a>	Multiple sequence alignment	(Edgar, 2004)
Conservation analysis			
ClustalX	<a href="http://www.ebi.ac.uk/Tools/clustalw2/index.html">http://www.ebi.ac.uk/Tools/clustalw2/index.html</a>	Conservation analysis and visualization	(Larkin et al., 2007)
ConSeq	<a href="http://conseq.tau.ac.il/">http://conseq.tau.ac.il/</a>	Conservation analysis and visualization	(Berezin et al., 2004)
ConSSeq	<a href="http://sms.cbi.cnpia.embrapa.br/SMS/STINGm/conseq/">http://sms.cbi.cnpia.embrapa.br/SMS/STINGm/conseq/</a>	Conservation analysis and visualization	(Higa et al., 2004)
ConSurf	<a href="http://consurf.tau.ac.il/">http://consurf.tau.ac.il/</a>	Conservation analysis and visualization	(Glaser et al., 2003; Landau et al., 2005)
Jalview	<a href="http://www.jalview.org/">http://www.jalview.org/</a>	MSA visualization	(Clamp et al., 2004)
MatrixPlot	<a href="http://www.cbs.dtu.dk/services/MatrixPlot/">http://www.cbs.dtu.dk/services/MatrixPlot/</a>	Conservation analysis and visualization	(Gorodkin et al., 1999)
MultiDisp	<a href="http://bioinf.uta.fi/cgi-bin/MultiDisp.cgi">http://bioinf.uta.fi/cgi-bin/MultiDisp.cgi</a>	Conservation analysis and visualization	(Riikonen and Vihinen, in preparation)
ProCon		Conservation analysis and visualization	(Shen and Vihinen, 2004)
Stability changes prediction			
Auto-Mute	<a href="http://proteins.gmu.edu/automute/AUTO-MUTE.html">http://proteins.gmu.edu/automute/AUTO-MUTE.html</a>	Stability changes prediction	(Masso and Vaisman, 2008)
CUPSAT	<a href="http://cupsat.tu-bs.de/">http://cupsat.tu-bs.de/</a>	Stability changes prediction	(Parthiban et al., 2006)
Dmutant	<a href="http://sparks.informatics.iupui.edu/hzhou/mutation.html">http://sparks.informatics.iupui.edu/hzhou/mutation.html</a>	Stability changes prediction	(Zhou and Zhou, 2002)
Eris	<a href="http://troll.med.unc.edu/eris/login.php">http://troll.med.unc.edu/eris/login.php</a>	Stability changes prediction	(Yin et al., 2007)
FoldX	<a href="http://foldx.crg.es/">http://foldx.crg.es/</a>	Folding and stability changes prediction	(Guerois et al., 2002)
I-Mutant 2.0	<a href="http://gpcr2.biocomp.unibo.it/cgi/predictors/I-Mutant2.0/I-Mutant2.0.cgi">http://gpcr2.biocomp.unibo.it/cgi/predictors/I-Mutant2.0/I-Mutant2.0.cgi</a>	Stability changes prediction	(Capriotti, et al., 2005a, b)
MuPRO	<a href="http://www.ics.uci.edu/%7Ebaldig/mutation.html">http://www.ics.uci.edu/%7Ebaldig/mutation.html</a>	Stability changes prediction	(Cheng et al., 2006)
PoPMuSiC	<a href="http://babylone.ulb.ac.be/popmusic/">http://babylone.ulb.ac.be/popmusic/</a>	Stability changes prediction	(Gilis and Rooman, 2000)
SCide	<a href="http://www.enzim.hu/scide/ide2.html">http://www.enzim.hu/scide/ide2.html</a>	Stability changes prediction	(Dosztányi et al., 2003)
SCpred	<a href="http://www.enzim.hu/scpred/pred.html">http://www.enzim.hu/scpred/pred.html</a>	Stability changes prediction	(Dosztányi et al., 1997)
SRide	<a href="http://sride.enzim.hu/">http://sride.enzim.hu/</a>	Stability changes prediction	(Magyar et al., 2005)
Disorder prediction			
CAST		Disorder prediction	(Promponas et al., 2000)
DisEMBL	<a href="http://dis.embl.de/">http://dis.embl.de/</a>	Disorder prediction	(Linding et al., 2003a)
Disopred	<a href="http://bioinf.cs.ucl.ac.uk/disopred/disopred.html">http://bioinf.cs.ucl.ac.uk/disopred/disopred.html</a>	Disorder prediction	(Ward et al., 2004)
DISpro	<a href="http://scratch.proteomics.ics.uci.edu/">http://scratch.proteomics.ics.uci.edu/</a>	Disorder prediction	(Cheng et al., 2005)
Disprot	<a href="http://www.ist.temple.edu/disprot/predictor.php">http://www.ist.temple.edu/disprot/predictor.php</a>	Disorder prediction	(Obradović et al., 2003; Peng et al., 2005; Vucetic et al., 2003)
DRIPPRED	<a href="http://www.sbc.su.se/~maccallr/disorder/">http://www.sbc.su.se/~maccallr/disorder/</a>	Disorder prediction	(Prilusky et al., 2005)
FoldIndex	<a href="http://bip.weizmann.ac.il/fldbin/index">http://bip.weizmann.ac.il/fldbin/index</a>	Prediction of folding	(Galitskaya et al., 2006)
FoldUnfold	<a href="http://skuld.protres.ru/~mlobanov/ogu/ogu.cgi">http://skuld.protres.ru/~mlobanov/ogu/ogu.cgi</a>	Disorder prediction	(Linding et al., 2003b)
GlobPlot	<a href="http://globplot.embl.de/">http://globplot.embl.de/</a>	Disorder prediction	(Su et al., 2007)
iPDA	<a href="http://biominer.bime.ntu.edu.tw/ipda/">http://biominer.bime.ntu.edu.tw/ipda/</a>	Disorder prediction	(Dosztányi et al., 2005)
IUPred	<a href="http://iupred.enzim.hu/">http://iupred.enzim.hu/</a>	Disorder prediction	(Liu and Rost, 2003)
NORSp	<a href="http://cubic.bioc.columbia.edu/services/NORSp/">http://cubic.bioc.columbia.edu/services/NORSp/</a>	Disorder prediction	(Obradović et al., 2005; Romero et al., 2001)
PONDR	<a href="http://www.pondr.com/">http://www.pondr.com/</a>	Disorder prediction	(Hirose et al., 2007; Shimizu et al., 2007a, b)
POODLE	<a href="http://mbs.cbrc.jp/poodle/poodle.html">http://mbs.cbrc.jp/poodle/poodle.html</a>	Disorder prediction	(Ishida and Kinoshita, 2007)
PrDOS	<a href="http://prdos.hgc.jp">http://prdos.hgc.jp</a>	Disorder prediction	(Coeytaux and Poupon, 2005)
PreLink	<a href="http://genomics.eu.org/spip/PreLink">http://genomics.eu.org/spip/PreLink</a>	Disorder prediction	(Yang et al., 2005)
RONN	<a href="http://www.strubi.ox.ac.uk/RONN">http://www.strubi.ox.ac.uk/RONN</a>	Disorder prediction	(Wootton, 1994)
SEG	<a href="http://mendel.imp.ac.at/METHODS/seg.server.html">http://mendel.imp.ac.at/METHODS/seg.server.html</a>	Disorder prediction	(Vullo et al., 2006)
Spritz	<a href="http://protein.cribi.unipd.it/spritz/">http://protein.cribi.unipd.it/spritz/</a>	Disorder prediction	
Analysis of interatomic contacts			
CMA	<a href="http://ligin.weizmann.ac.il/cma/">http://ligin.weizmann.ac.il/cma/</a>	Analysis of interatomic contacts	(Sobolev et al., 2005)
CSU	<a href="http://bip.weizmann.ac.il/oca-bin/lpccsu">http://bip.weizmann.ac.il/oca-bin/lpccsu</a>	Analysis of interatomic contacts	(Sobolev et al., 1999)
KiNG	<a href="http://kinemage.biochem.duke.edu/software/king.php">http://kinemage.biochem.duke.edu/software/king.php</a>	Molecular graphics	(Lovell et al., 2003)
MolProbity	<a href="http://molprobity.biochem.duke.edu/">http://molprobity.biochem.duke.edu/</a>	Analysis of interatomic contacts and packing, structure validation	(Davis et al., 2004)

**Table 1. Continued**

Service name	URL	Description	Reference
PROBE	<a href="http://kinemage.biochem.duke.edu/software/probe.php">http://kinemage.biochem.duke.edu/software/probe.php</a>	Analysis of interatomic contacts and packing	(Word et al., 2000; Word et al., 1999)
PyMOL	<a href="http://pymol.sourceforge.net/">http://pymol.sourceforge.net/</a>	Molecular graphics	(DeLano, 2002)
RankViaContact	<a href="http://bioinf.uta.fi/RankViaContact.html">http://bioinf.uta.fi/RankViaContact.html</a>	Analysis and visualization of interatomic contacts	(Shen and Vihinen, 2003)
Aggregation prediction			
Aggrescan	<a href="http://bioinf.uab.es/aggrescan/">http://bioinf.uab.es/aggrescan/</a>	Aggregation prediction	(Conchillo-Sole et al., 2007)
PASTA	<a href="http://protein.cribi.unipd.it/pasta/">http://protein.cribi.unipd.it/pasta/</a>	Aggregation prediction	(Trovato et al., 2007)
TANGO	<a href="http://tango.embl.de/">http://tango.embl.de/</a>	Aggregation prediction	(Fernandez-Escamilla et al., 2004)
Waltz	<a href="http://switpc7.vub.ac.be/cgi-bin/submit.cgi">http://switpc7.vub.ac.be/cgi-bin/submit.cgi</a>	Aggregation prediction	(Maurer-Stroh et al., submitted for publication)
Other			
ExPASy	<a href="http://ca.expasy.org/tools/#ptm">http://ca.expasy.org/tools/#ptm</a>	Posttranslational modification prediction tools	
Proteomics tools			
SABLE	<a href="http://sable.cchmc.org/">http://sable.cchmc.org/</a>	Prediction of solvent accessibilities, 2D structures and transmembrane domains	(Adamczak et al., 2004, 2005; Wagner et al., 2005)
SNPeffect	<a href="http://snpeffect.vib.be">http://snpeffect.vib.be</a>	Prediction platform (metaserver) and database	(Reumers et al., 2006)

the protein may become vulnerable to protease digestion. Disorder is further discussed in the reviews by Bourhis et al. [2007], Dosztányi et al. [2007], and Ferron et al. [2006].

The methods predicting protein structural disorder are based on protein amino acid composition as well as energy profiles and physicochemical properties of the amino acids, specific sequence patterns, missing X-ray coordinates, and B-factors. A number of disorder prediction methods are based on machine learning methods, such as support vector machines (SVM) and self-organizing maps (SOM). As no clear definition of the concept of disorder exists, the different methods predict disorder by varying means. It should be noted that the methods discussed here have not been developed for the study of the effects of missense mutations but, according to our experience, they can be used for that purpose with certain reservations. Given that several of these methods would predict a mutation to increase or decrease the disordered structure content of a protein, one could conclude that the mutation probably has damaging effects on the structure and thereby function of the protein.

Several attempts have been made to build disorder predictors that would operate solely on sequence data. These methods, for example, SEG [Wootton, 1994] and CAST [Promponas et al., 2000], divide sequences into regions of low or high complexity. Low-complexity regions are compositionally biased regions that are rarely defined in protein 3D structures [Saqi and Sternberg, 1994]. SEG and CAST mainly detect repetitive segments in sequences, which often exhibit structural disorder. However, not all regions with low sequence complexity are disordered, and vice versa [Romero et al., 2001]. Other prediction methods operating on sequence information, PONDR [Obradović et al., 2005; Romero et al., 2001], iPDA [Su et al., 2007], and POODLE-L [Hirose et al., 2007], analyze disorder propensities based on amino acid properties and neural networks (NNs) (PONDR), radial basis function networks (iPDA), or SVMs (POODLE-L, Spritz) [Vullo et al., 2006], that have been trained on a set of disordered and ordered sequences. PreLink assigns probabilities for amino acid residues to occur in disordered regions combined with the distance of each amino acid from the nearest hydrophobic cluster [Coeytaux and Poupon, 2005]. Globplot is a tool for recognizing globular and disordered regions within amino acid sequences based on Russell/Linding secondary structure-forming propensities [Linding et al., 2003b]. Another method using secondary structure-forming capacity as a parameter is NORSp, which estimates the secondary structure content of the amino acid sequence, and assigns those sequence segments with no predicted

2D structure as disordered [Liu and Rost, 2003]. IUPred [Dosztányi et al., 2005a] estimates the capacity of polypeptides to form stabilizing interresidue contacts based on amino acid chemical types and their sequence environment. The sequence regions with less contact-forming capacity are defined as disordered [Dosztányi et al., 2005b]. FoldUnfold utilizes expected packing densities for amino acid sequences [Galzitskaya et al., 2006a], such that weak expected packing densities point to disordered regions [Galzitskaya et al., 2006b].

RONN predicts disorder by comparing the input sequence to other sequences of known folding state, and the alignment scores against these sequences are used to classify the input sequence as ordered or disordered using a neural network [Yang et al., 2005]. The PrDOS method [Ishida and Kinoshita, 2007] utilizes template proteins (assuming that disorder is conserved in protein families) complementing the amino acid sequence profile generated by a SVM.

In the DRIP-PRED method [MacCallum, 2004], self-organizing maps have been trained on protein sequences with known structure. The target sequence windows are mapped onto the SOM, and when sequence windows map onto regions not well represented in the PDB, those sequences are predicted to be disordered. This approach may be problematic because the PDB is biased and does not contain all types of structures.

The methods POODLE-S [Shimizu et al., 2007a], DisPRO [Cheng et al., 2005], DISOPRED2 [Ward et al., 2004], and DisEMBL [Linding et al., 2003a], are NN-based methods that define disorder as missing coordinates in high-resolution X-ray crystal structure electron density maps. The DisEMBL method requires that the disordered regions must reside within loops or coils, and both POODLE-S and DisEMBL also take B-factors into account so that highly dynamic loops are considered to be disordered.

The regions lacking coordinates in crystal structures are commonly classified as disordered both in the prediction methods and in experiments assessing the reliability of the methods, such as in the Critical Assessment of Techniques for Protein Structure Prediction (CASP) [Bordoli et al., 2007; Jin and Dunbrack, 2005]. However, missing electron density is not a perfect definition of disorder, because crystallization may impose order on regions that would be disordered in solution, and conversely, missing electron density may not necessarily prove the lack of ordered structure. Some regions may be disordered with respect to the rest of the structure in a crystal, although they may be internally ordered [Jin and Dunbrack, 2005]. The disadvantage in using B-factors in disorder prediction is that they can vary greatly within a single



structure as a result of the effects of local packing [Smith et al., 2003] and, for example, a residue side chain may have alternative conformations leading into an elevated B-factor that does not indicate disorder (Fig. 2F).

## Aggregation

An increased level of  $\beta$ -structure is characteristic of different types of protein aggregates, such as amyloid fibrils and amorphous aggregates [Jiménez et al., 1999; Ohnishi and Takano, 2004; Rousseau et al., 2006]. In addition to those proteins involved in amyloid diseases (which include Alzheimer Disease, Parkinson Disease, and type II diabetes, as well as the spongiform encephalopathies), it has been shown that diverse proteins not related to amyloid disease can aggregate under destabilizing conditions [Chiti et al., 1999; Fandrich et al., 2001; Guijarro et al., 1998], and that normal proteins can become toxic upon fibrillation [Bucciantini et al., 2004].

Missense mutations can change the properties of a protein so that its tendency to aggregate increases. It has been suggested that the composition and the primary structure of a protein determine to a large extent its propensity to aggregate, and even small alterations may have a considerable effect in the solubility of the protein. Aggregation has been shown to be modulated by very short stretches of specific amino acids that can act as facilitators or inhibitors of amyloid fibril formation [Ivanova et al., 2004; Ventura et al., 2004].

A number of algorithms have been developed for the prediction of aggregation propensities of proteins [Chiti et al., 2003; DuBay et al., 2004; Tartaglia et al., 2005; Thompson et al., 2006; Yoon and Welsh, 2004]. The following methods are also available as Web services: The AGGRESCAN [Conchillo-Solé et al., 2007] method is based on aggregation propensity values assigned to each amino acid residue determined by experimental studies [de Groot et al., 2006]. TANGO [Fernandez-Escamilla et al., 2004] is a method based on secondary structure propensities and estimation of desolvation energy. PASTA [Trovato et al., 2007] is based on sequence-specific interaction energies between pairs of protein fragments calculated from statistical analysis of the native folds of globular proteins [Trovato et al., 2006].

The methods for the prediction of  $\beta$ -aggregation are mostly based on physicochemical properties of the input sequences. They are relatively straightforward because of the regular structural arrangement and the important role of side chains in  $\beta$ -sheet aggregates [Azriel and Gazit, 2001; Gazit, 2002; Gsponer et al., 2003; López de la Paz and Serrano, 2004; Williams et al., 2006].

## Structural Considerations

When a residue is replaced by another residue in a missense mutation, many of its chemical and physical properties may be altered (Fig. 1). The substitution may cause major structural arrangements, especially when the wild-type residue is smaller than the substituting one. Whether the new side chain can be fitted into the structure without major structural rearrangements, and how this can be achieved, can be studied by rotamer analysis. The new side chain is modeled into the structure by, for example, PyMOL [DeLano, 2002], KiNG [Lovell et al., 2003], Discovery Studio (Accelrys, San Diego, CA), or Swiss-PDB-Viewer [Guex and Peitsch, 1997], and hydrogens are added to the structure by, for example, Reduce [Word et al., 1999]. Overpacking can be measured by rotating each of the mutated side chains over full range of side chain  $\chi$  angles (Fig. 2E). Only the substituted side chain is allowed to move during the analyses. The rotatable side chain is created and an automated sampling of torsional

angles is done with, for example, the Autobondrot procedure under PROBE [Word et al., 1999, 2000]. The acceptable conformations for a mutated side chain have a total score of above  $-1.0$ , allowing for small local perturbations to take place in the structure [Lovell et al., 2000]. A lower score indicates that the side chain does not fit into the structure in any conformation without deleterious changes in the protein scaffolding. The highest scoring rotamers are then selected and modeled into the structure for further analysis (Fig. 2D). The created structures can be verified by MolProbity [Davis et al., 2007], a Web server providing all-atom contact analysis as well as Ramachandran and rotamer distributions. The quality of the structure can be studied by the protein structure verification tools PROCHECK [Morris et al., 1992] or WHAT\_CHECK [Hooft et al., 1996]. When available, experimentally solved structures are used as templates in the analysis of structural effects caused by mutations. Protein structure prediction and molecular modeling can provide valuable information when the 3D structure of the protein of interest has not been determined [Baker and Sali, 2001]. Structural and biological/medical interpretations can also be quite accurate when based on modeled protein structures [Khan and Vihinen, submitted].

## Residue Contacts and Stability

Compromised folding and decreased stability of the protein product are the major molecular pathogenic consequences of a missense mutation [Bross et al., 1999; Wang and Moulton, 2001; Yue et al., 2005]. Protein folding and stability are closely coupled and, for disease mutants, folding can be slowed so much that most molecules are targeted for recycling by the quality control machinery in the endoplasmic reticulum [Plempner and Wolf, 1999]. Alternatively, the protein fails to fold correctly as a result of a mutation, which may have a detrimental impact on protein function.

Missense mutations may have an effect on the stability of the protein via overpacking (Fig. 2D), altered contacts between amino acid side chains, reduction in hydrophobic area, altered structural strain in the protein backbone introduced by proline residues, or changes in electrostatics. These alterations may have an effect on the free energy difference between the folded and unfolded states of the protein by causing changes in interaction energy between amino acids, or affecting the entropy of the system or local rigidity of the structure [Yue et al., 2005].

Chemical bonds and interactions between amino acid side chains determine the two- and three-dimensional fold and detailed shape of a protein. Hydrophobic interactions in the protein core are crucial in maintaining the overall structural stability of the protein, and introducing a charged residue into the core generally destabilizes the protein [Chasman and Adams, 2001]. The net effect of a number of hydrophobic interactions determines the stability of the protein core, and even the more subtle alterations in these interactions could have a detrimental effect on the structural integrity of a protein [Matthews 1995; Sandberg et al., 1995; Serrano et al., 1992; Shortle et al., 1990]. The vulnerability of the hydrophobic core is illustrated by the fact that the probability of a mutation to be pathogenic increases with a decrease in the solvent accessibility of the site [Vitkup et al., 2003]. The interactions between side chains on the surface of a protein define and maintain local structure, the details of which may be crucial for ligand or substrate binding or for interactions with partner proteins or DNA.

After modeling the mutated side chain into the structure, its effect on the chemical bonds with neighboring residues and changes in the solvent accessible surface of the residue atoms can be studied by the CSU service [Sobolev et al., 1999], or visually by the MAGE/

PROBE system [Word et al., 2000], KiNG [Lovell et al., 2003], or molecular modeling software packages. RankViaContact is a service for calculation of residue–residue contact energies [Shen and Vihinen, 2003]. Strong contacts are favorable for stability, while weaker contacts between residues may point to functional regions [Beadle and Shoichet, 2002]. The effects of mutations on contact energies can provide insight into the structure–function relationships of the mutated positions at the protein level.

There are several services available for the prediction of the effects of mutations on protein stability. Cupsat [Parthiban et al., 2006], Eris [Yin et al., 2007], FoldX [Schymkowitz et al., 2005], DMUTANT [Zhou and Zhou, 2002], and PoPMuSic [Gilis and Rooman, 2000] calculate mutational free energy changes of the protein based on its 3D structure. I-Mutant 2.0 [Capriotti et al., 2005a, b], MuPro [Cheng et al., 2006], and the method developed by Shen et al. [2008] utilize support vector machines or neural networks to predict the effect of the substitution on protein stability. Auto-Mute [Masso and Vaisman, 2008] is a method that combines a knowledge-based statistical potential with machine learning techniques in the prediction. SRide [Magyar et al., 2005] and SCide [Dosztányi et al., 2003a] predict stabilizing residues based on long-range interactions in protein structures. SRide includes hydrophobicity and conservation of residues as additional parameters. SCPred is a method based on differences in sequential neighborhood [Dosztányi et al., 2003b].

## Electrostatics

Patches of electrostatic potential are often indicators of a binding surface, usually to a molecule with a potential of opposite sign [Honig and Nicholls, 1995]. However, this is not always the case. Some interfaces exploit electrostatic interactions to drive binding, while in others hydrophobic residues appear to be the dominant surface feature [Sheinerman and Honig, 2002]. Surface charge–charge relationships are also important in maintaining the stability of the protein [Strickler et al., 2006]. Changes in electrostatic potential affect the properties of proteins in many ways. Mutations that induce local changes in electrostatic surface potential may have a crucial effect on ligand binding or specificity, and electrostatic alterations may affect protein folding and stability. Qualitative measures of electrostatic surface potentials can be calculated, for example, with PyMOL [DeLano, 2002] or Delphi [Rocchia et al., 2002].

## Pathogenic-or-Not Predictors

Several prediction methods that aim at sorting mutations according to their pathogenicity, such as SIFT [Ng and Henikoff, 2001] and MAPP [Stone and Sidow, 2005], are based on phylogenetic information, mainly assuming that the majority of substitutions observed between humans and closely related species are functionally neutral. The PhD-SNP method [Capriotti et al., 2006] utilizes SVM classifiers based on sequence environment and conservation. It has been shown that combining information obtained from the multiple sequence alignment with structural information can increase the prediction accuracy [Saunders and Baker, 2002]. Some methods, for example, nsSNPAnalyzer [Bao et al., 2005], PolyPhen [Ramensky et al., 2002], and SNPs3D [Yue et al., 2006], combine available structural information with the multiple sequence alignments to reach more accurate results. Align-GVGD [Mathe et al., 2006] and SNAP [Bromberg and Rost, 2007] combine information about the biochemical properties of the wild-type and the substituting residue with evolutionary information.

Some methods use structural and functional annotation from the Swiss-Prot database in addition to structure and sequence modelling [Ferrer-Costa et al., 2002, 2004; Sunyaev et al., 2000, 2001b; Wang and Moulton, 2001]. The functional annotation is used to identify the residues that are part of a binding site, active site, or disulfide bond. It is presumed that changes at these positions would have a major effect on protein function.

These prediction methods can be useful, in addition to their obvious function of predicting whether a mutation is pathogenic, in deducing the mechanism by which a mutation causes a disease. Indeed, some of these methods may predict a known pathogenic mutation to be benign, but this information can be valuable in ruling out some possible disease mechanisms.

## PON-P: Pathogenic-or-Not Pipeline

We are currently developing a service providing simultaneous access to the numerous prediction methods described in this paper. When studying the effects of mutations by bioinformatics methods, submitting sequence and mutation data to the various predictors requires a considerable amount of work and time, especially when the number of mutations in a given sequence is large. A service that simultaneously submits the input data provided by the user to selected prediction methods, as well as parses the outputs of individual methods into a single output, will simplify the process and provide results faster and more conveniently. PON-P—the Pathogenic-or-Not Pipeline (Fig. 2A)—will initially feature all the pathogenic-or-not predictors described in the previous chapters, as well as links and descriptions for all prediction methods described in this article. In the near future there will be a user-friendly submission form for analyses of different kinds of mutations. PON-P is currently being developed to contain all the available predictors for disorder, aggregation, tolerance, and stability. The Pipeline will be freely available at <http://bioinf.uta.fi/PON-P>.

## Conclusion

As the number of known variants in the human genome increases, the determination of positions likely to be disease-associated has become an important and challenging problem. There are numerous bioinformatics methods available for the analysis of the molecular consequences of missense mutations. Several of the methods are very specific, and dedicated to the analysis of a single feature. However, they may analyze the same property from different points of view. For example, structural changes may originate from changes in side-chain size, hydrophobicity, altered contact-forming properties, aggregation, or introduced disorder. To make sophisticated choices of the most suitable prediction methods and to be able to interpret the results correctly, it is of utmost importance to be familiar with the theory and limitations of the various methods. The Pathogenic-or-Not Pipeline (PON-P) is a service providing access to various mutation analysis methods, facilitating their use.

## Acknowledgments

We thank Kathryn Rannikko for language revision.

## References

- Adamczak R, Porollo A, Meller J. 2004. Accurate prediction of solvent accessibility using neural networks-based regression. *Proteins* 56:753–767.



- Adamczak R, Porollo A, Meller J. 2005. Combining prediction of secondary structure and solvent accessibility in proteins. *Proteins* 59:467–475.
- Ahola V, Aittokallio T, Uusipaikka E, Vihinen M. 2004. Statistical methods for identifying conserved residues in multiple sequence alignment. *Stat Appl Genet Mol Biol* 3:28.
- Ahola V, Aittokallio T, Vihinen M, Uusipaikka E. 2006. A statistical score for assessing the quality of multiple sequence alignments. *BMC Bioinformatics* 7:484.
- Azriel R, Gazit E. 2001. Analysis of the minimal amyloid-forming fragment of the islet amyloid polypeptide. An experimental support for the key role of the phenylalanine residue in amyloid formation. *J Biol Chem* 276:34156–34161.
- Baker D, Sali A. 2001. Protein structure prediction and structural genomics. *Science* 294:93–96.
- Bao L, Zhou M, Cui Y. 2005. nsSNPAnalyzer: identifying disease-associated nonsynonymous single nucleotide polymorphisms. *Nucleic Acids Res* 33:W480–W482.
- Beadle BM, Shoichet BK. 2002. Structural bases of stability–function tradeoffs in enzymes. *J Mol Biol* 321:285–296.
- Berezin C, Glaser F, Rosenberg J, Paz I, Pupko T, Fariselli P, Casadio R, Ben-Tal N. 2004. ConSeq: the identification of functionally and structurally important residues in protein sequences. *Bioinformatics* 20:1322–1324.
- Bordoli L, Kiefer F, Schwede T. 2007. Assessment of disorder predictions in CASP7. *Proteins* 69(Suppl 8):129–136.
- Bourhis JM, Canard B, Longhi S. 2007. Predicting protein disorder and induced folding: from theoretical principles to practical applications. *Curr Protein Peptide Sci* 8:135–149.
- Briscoe AD, Gaur C, Kumar S. 2004. The spectrum of human rhodopsin disease mutations through the lens of interspecific variation. *Gene* 332:107–118.
- Bromberg Y, Rost B. 2007. SNAP: predict effect of non-synonymous polymorphisms on function. *Nucleic Acids Res* 35:3823–3835.
- Bross P, Corydon TJ, Andresen BS, Jorgensen MM, Bolund L, Gregersen N. 1999. Protein misfolding and degradation in genetic diseases. *Hum Mutat* 14:186–198.
- Bucciantini M, Calloni G, Chiti F, Formigli L, Nosi D, Dobson CM, Stefani M. 2004. Prefibrillar amyloid protein aggregates share common features of cytotoxicity. *J Biol Chem* 279:31374–31382.
- Burke DF, Worth CL, Priego EM, Cheng T, Smink LJ, Todd JA, Blundell TL. 2007. Genome bioinformatic analysis of nonsynonymous SNPs. *BMC Bioinformatics* 8:301.
- Capriotti E, Calabrese R, Casadio R. 2006. Predicting the insurgence of human genetic diseases associated to single point protein mutations with support vector machines and evolutionary information. *Bioinformatics* 22:2729–2734.
- Capriotti E, Fariselli P, Calabrese R, Casadio R. 2005a. Predicting protein stability changes from sequences using support vector machines. *Bioinformatics* 21(Suppl 2):ii54–ii58.
- Capriotti E, Fariselli P, Casadio R. 2005b. I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure. *Nucleic Acids Res* 33:W306–W310.
- Cargill M, Altshuler D, Ireland J, Sklar P, Ardlie K, Patil N, Shaw N, Lane CR, Lim EP, Kalyanaraman N, Nemesh J, Ziaugra L, Friedland L, Rolfe A, Warrington J, Lipshutz R, Daley GQ, Lander ES. 1999. Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nat Genet* 22:231–238.
- Cavallo A, Martin AC. 2005. Mapping SNPs to protein sequence and structure data. *Bioinformatics* 21:1443–1450.
- Chasman D, Adams RM. 2001. Predicting the functional consequences of non-synonymous single nucleotide polymorphisms: structure-based assessment of amino acid variation. *J Mol Biol* 307:683–706.
- Cheng J, Randall A, Baldi P. 2006. Prediction of protein stability changes for single-site mutations using support vector machines. *Proteins* 62:1125–1132.
- Cheng J, Randall AZ, Sweredoski MJ, Baldi P. 2005. SCRATCH: a protein structure and structural feature prediction server. *Nucleic Acids Res* 33:W72–W76.
- Chiti F, Stefani M, Taddei N, Ramponi G, Dobson CM. 2003. Rationalization of the effects of mutations on peptide and protein aggregation rates. *Nature* 424:805–808.
- Chiti F, Webster P, Taddei N, Clark A, Stefani M, Ramponi G, Dobson CM. 1999. Designing conditions for in vitro formation of amyloid protofilaments and fibrils. *Proc Natl Acad Sci USA* 96:3590–3594.
- Clamp M, Cuff J, Searle SM, Barton GJ. 2004. The Jalview Java alignment editor. *Bioinformatics* 20:426–427.
- Coeytaux K, Poupon A. 2005. Prediction of unfolded segments in a protein sequence based on amino acid composition. *Bioinformatics* 21:1891–1900.
- Conchillo-Solé O, de Groot NS, Aviles FX, Vendrell J, Daura X, Ventura S. 2007. AGGRESKAN: a server for the prediction and evaluation of “hot spots” of aggregation in polypeptides. *BMC Bioinformatics* 8:65.
- Cotton RG, Auerbach AD, Beckmann JS, Blumenfeld OO, Brookes AJ, Brown AF, Carrera P, Cox DW, Gottlieb B, Greenblatt MS, Hilbert P, Lehtväliho H, Liang P, Marsh S, Nebert DW, Povey S, Rossetti S, Scriver CR, Summar M, Tolan DR, Verma IC, Vihinen M, den Dunnen JT. 2008. Recommendations for locus-specific databases and their curation. *Hum Mutat* 29:2–5.
- Dantzer J, Moad C, Heiland R, Mooney S. 2005. MutDB services: interactive structural analysis of mutation data. *Nucleic Acids Res* 33:W311–W314.
- Davis IW, Leaver-Fay A, Chen VB, Block JN, Kapral GJ, Wang X, Murray LW, Arendall III WB, Snoeyink J, Richardson JS, Richardson DC. 2007. MolProbity: all-atom contacts and structure validation for proteins and nucleic acids. *Nucleic Acids Res* 35:W375–W383.
- de Groot NS, Aviles FX, Vendrell J, Ventura S. 2006. Mutagenesis of the central hydrophobic cluster in A $\beta$ 42 Alzheimer’s peptide. Side-chain properties correlate with aggregation propensities. *FEBS J* 273:658–668.
- DeLano W. 2002. The PyMOL molecular graphics system. Palo Alto, CA: DeLano Scientific.
- den Dunnen JT, Antonarakis SE. 2000. Mutation nomenclature extensions and suggestions to describe complex mutations: a discussion. *Hum Mutat* 15:7–12.
- Do CB, Mahabhashyam MS, Brudno M, Batzoglou S. 2005. ProbCons: probabilistic consistency-based multiple sequence alignment. *Genome Res* 15:330–340.
- Dosztányi Z, Csizsók V, Tompa P, Simon I. 2005a. IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics* 21:3433–3434.
- Dosztányi Z, Csizsók V, Tompa P, Simon I. 2005b. The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. *J Mol Biol* 347:827–839.
- Dosztányi Z, Fiser A, Simon I. 1997. Stabilization centers in proteins: identification, characterization and predictions. *J Mol Biol* 272:597–612.
- Dosztányi Z, Magyar C, Tusnády G, Simon I. 2003a. SCide: identification of stabilization centers in proteins. *Bioinformatics* 19:899–900.
- Dosztányi Z, Magyar C, Tusnády GE, Cserzo M, Fiser A, Simon I. 2003b. Servers for sequence–structure relationship analysis and prediction. *Nucleic Acids Res* 31:3359–3363.
- Dosztányi Z, Sándor M, Tompa P, Simon I. 2007. Prediction of protein disorder at the domain level. *Curr Protein Pept Sci* 8:161–171.
- DuBay KF, Pawar AP, Chiti F, Zurdo J, Dobson CM, Vendruscolo M. 2004. Prediction of the absolute aggregation rates of amyloidogenic polypeptide chains. *J Mol Biol* 341:1317–1326.
- Dunker AK, Brown CJ, Lawson JD, Iakoucheva LM, Obradović Z. 2002. Intrinsic disorder and protein function. *Biochemistry* 41:6573–6582.
- Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 32:1792–1797.
- Emanuelsson O, Brunak S, von Heijne G, Nielsen H. 2007. Locating proteins in the cell using TargetP, SignalP and related tools. *Nat Protoc* 2:953–971.
- Fandrich M, Fletcher MA, Dobson CM. 2001. Amyloid fibrils from muscle myoglobin. *Nature* 410:165–166.
- Fernandez-Escamilla AM, Rousseau F, Schymkowitz J, Serrano L. 2004. Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins. *Nat Biotechnol* 22:1302–1306.
- Ferrer-Costa C, Gelpí JL, Zamakola L, Parraga I, de la Cruz X, Orozco M. 2005. PMUT: a web-based tool for the annotation of pathological mutations on proteins. *Bioinformatics* 21:3176–3178.
- Ferrer-Costa C, Orozco M, de la Cruz X. 2002. Characterization of disease-associated single amino acid polymorphisms in terms of sequence and structure properties. *J Mol Biol* 315:771–786.
- Ferrer-Costa C, Orozco M, de la Cruz X. 2004. Sequence-based prediction of pathological mutations. *Proteins* 57:811–819.
- Ferron F, Longhi S, Canard B, Karlin D. 2006. A practical overview of protein disorder prediction methods. *Proteins* 65:1–14.
- Finn RD, Tate J, Mistry J, Coghill PC, Sammut SJ, Hotz HR, Ceric G, Forslund K, Eddy SR, Sonnhammer EL, Bateman A. 2008. The Pfam protein families database. *Nucleic Acids Res* 36:D281–D288.
- Galzitskaya OV, Garbuzynskiy SO, Lobanov MY. 2006a. FoldUnfold: web server for the prediction of disordered regions in protein chain. *Bioinformatics* 22:2948–2949.
- Galzitskaya OV, Garbuzynskiy SO, Lobanov MY. 2006b. Prediction of amyloidogenic and disordered regions in protein chains. *PLoS Comput Biol* 2:e177.
- Gazit E. 2002. A possible role for  $\pi$ -stacking in the self-assembly of amyloid fibrils. *FASEB J* 16:77–83.
- Giardine B, Riemer C, Hefferon T, Thomas D, Hsu F, Zielinski J, Sang Y, Elnitski L, Cutting G, Trumbower H, Kern A, Kuhn R, Patrinos GP, Hughes J, Higgs D, Chui D, Scriver C, Phommavanh M, Patnaik SK, Blumenfeld O, Gottlieb B, Vihinen M, Väliaho J, Kent J, Miller W, Hardison RC. 2007. PhenCode: connecting ENCODE data with mutations and phenotype. *Hum Mutat* 28:554–562.
- Gilis D, Rooman M. 2000. PoPMuSiC, an algorithm for predicting protein mutant stability changes: application to prion proteins. *Protein Eng* 13:849–856.

- Glaser F, Pupko T, Paz I, Bell RE, Bechor-Shental D, Martz E, Ben-Tal N. 2003. ConSurf: identification of functional regions in proteins by surface-mapping of phylogenetic information. *Bioinformatics* 19:163–164.
- Gloor GB, Martin LC, Wahl LM, Dunn SD. 2005. Mutual information in protein multiple sequence alignments reveals two classes of coevolving positions. *Biochemistry* 44:7156–7165.
- Golubchik T, Wise MJ, Eastel S, Jermini LS. 2007. Mind the gaps: evidence of bias in estimates of multiple sequence alignments. *Mol Biol Evol* 24:2433–2442.
- Gorodkin J, Starfjeldt HH, Lund O, Brunak S. 1999. MatrixPlot: visualizing sequence constraints. *Bioinformatics* 15:769–770.
- Gspöner J, Habberthür U, Caffisch A. 2003. The role of side-chain interactions in the early steps of aggregation: molecular dynamics simulations of an amyloid-forming peptide from the yeast prion Sup35. *Proc Natl Acad Sci USA* 100:5154–5159.
- Gueriois R, Nielsen JE, Serrano L. 2002. Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. *J Mol Biol* 320:369–387.
- Guex N, Peitsch MC. 1997. SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. *Electrophoresis* 18:2714–2723.
- Guijarro JI, Sunde M, Jones JA, Campbell ID, Dobson CM. 1998. Amyloid fibril formation by an SH3 domain. *Proc Natl Acad Sci USA* 95:4224–4228.
- Halushka MK, Fan JB, Bentley K, Hsie L, Shen N, Weder A, Cooper R, Lipshutz R, Chakravarti A. 1999. Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis. *Nat Genet* 22:239–247.
- Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA. 2005. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res* 33:D514–D517.
- Higa RH, Montagner AJ, Togawa RC, Kuser PR, Yamagishi ME, Mancini AL, Pappas Jr G, Miura RT, Horita LG, Neshich G. 2004. ConSseq: a web-based application for analysis of amino acid conservation based on HSSP database and within context of structure. *Bioinformatics* 20:1983–1985.
- Hirose S, Shimizu K, Kanai S, Kuroda Y, Noguchi T. 2007. POODLE-L: a two-level SVM prediction system for reliably predicting long disordered regions. *Bioinformatics* 23:2046–2053.
- Honig B, Nicholls A. 1995. Classical electrostatics in biology and chemistry. *Science* 268:1144–1149.
- Hooft RW, Vriend G, Sander C, Abola EE. 1996. Errors in protein structures. *Nature* 381:272.
- Hyytiäinen ER, Haapala K, Thompson J, Lappalainen I, Roiha M, Rantala I, Helin HJ, Jänne OA, Vihinen M, Palvimo JJ, Koivisto PA. 2002. Pattern of somatic androgen receptor gene mutations in patients with hormone-refractory prostate cancer. *Lab Invest* 82:1591–1598.
- Ishida T, Kinoshita K. 2007. PrDOS: prediction of disordered protein regions from amino acid sequence. *Nucleic Acids Res* 35:W460–W464.
- Ivanova MI, Sawaya MR, Gingery M, Attinger A, Eisenberg D. 2004. An amyloid-forming segment of  $\beta$ 2-microglobulin suggests a molecular model for the fibril. *Proc Natl Acad Sci USA* 101:10584–10589.
- Jiménez JL, Guijarro JI, Orlova E, Zurdo J, Dobson CM, Sunde M, Saibil HR. 1999. Cryo-electron microscopy structure of an SH3 amyloid fibril and model of the molecular packing. *EMBO J* 18:815–821.
- Jin Y, Dunbrack Jr RL. 2005. Assessment of disorder predictions in CASP6. *Proteins* 61(Suppl 7):167–175.
- Katoh K, Kuma K, Toh H, Miyata T. 2005. MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res* 33:511–518.
- Katoh K, Misawa K, Kuma K, Miyata T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res* 30:3059–3066.
- Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. 2002. The human genome browser at UCSC. *Genome Res* 12:996–1006.
- Landau M, Mayrose I, Rosenberg Y, Glaser F, Martz E, Pupko T, Ben-Tal N. 2005. ConSurf 2005: the projection of evolutionary conservation scores of residues on protein structures. *Nucleic Acids Res* 33:W299–W302.
- Lappalainen I, Giliani S, Franceschini R, Bonnefoy JY, Duckett C, Notarangelo LD, Vihinen M. 2000. Structural basis for SH2D1A mutations in X-linked lymphoproliferative disease. *Biochem Biophys Res Commun* 269:124–130.
- Lappalainen I, Thusberg J, Shen B, Vihinen M. 2008. Genome wide analysis of pathogenic SH2 domain mutations. *Proteins* 72:779–792.
- Lappalainen I, Vihinen M. 2002. Structural basis of ICF-causing mutations in the methyltransferase domain of DNMT3B. *Protein Eng* 15:1005–1014.
- Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ, Higgins DG. 2007. Clustal W and Clustal X version 2.0. *Bioinformatics* 23:2947–2948.
- Lau AY, Chasman DI. 2004. Functional classification of proteins and protein variants. *Proc Natl Acad Sci USA* 101:6576–6581.
- Lichtarge O, Bourne HR, Cohen FE. 1996. An evolutionary trace method defines binding surfaces common to protein families. *J Mol Biol* 257:342–358.
- Linding R, Jensen LJ, Diella F, Bork P, Gibson TJ, Russell RB. 2003a. Protein disorder prediction: implications for structural proteomics. *Structure* 11:1453–1459.
- Linding R, Russell RB, Neduva V, Gibson TJ. 2003b. GlobPlot: exploring protein sequences for globularity and disorder. *Nucleic Acids Res* 31:3701–3708.
- Liu J, Rost B. 2003. NORSp: predictions of long regions without regular secondary structure. *Nucleic Acids Res* 31:3833–3835.
- Livingston RJ, von Niederhausern A, Jegga AG, Crawford DC, Carlson CS, Rieder MJ, Gowrisankar S, Aronow BJ, Weiss RB, Nickerson DA. 2004. Pattern of sequence variation across 213 environmental response genes. *Genome Res* 14:1821–1831.
- Lockless SW, Ranganathan R. 1999. Evolutionarily conserved pathways of energetic connectivity in protein families. *Science* 286:295–299.
- Lopez de la Paz M, Serrano L. 2004. Sequence determinants of amyloid fibril formation. *Proc Natl Acad Sci USA* 101:87–92.
- Lovell SC, Davis IW, Arendall III WB, de Bakker PI, Word JM, Prisant MG, Richardson JS, Richardson DC. 2003. Structure validation by  $\alpha$  geometry:  $\phi$ ,  $\psi$  and  $C\beta$  deviation. *Proteins* 50:437–450.
- Lovell SC, Word JM, Richardson JS, Richardson DC. 2000. The penultimate rotamer library. *Proteins* 40:389–408.
- MacCallum R. 2004. Order/disorder prediction with self organizing maps. CASP6 Online Paper. <http://www.forcas.org/paper2127.html>
- Magyar C, Gromiha MM, Pujadas G, Tusnády GE, Simon I. 2005. SRide: a server for identifying stabilizing residues in proteins. *Nucleic Acids Res* 33:W303–W305.
- Marcotte EM, Pellegrini M, Ng HL, Rice DW, Yeates TO, Eisenberg D. 1999. Detecting protein function and protein-protein interactions from genome sequences. *Science* 285:751–753.
- Masso M, Vaisman II. 2008. Accurate prediction of stability changes in protein mutants by combining machine learning with structure based computational mutagenesis. *Bioinformatics* 24:2002–2009.
- Mathe E, Olivier M, Kato S, Ishioka C, Hainaut P, Tavtigian SV. 2006. Computational approaches for predicting the biological effect of p53 missense mutations: a comparison of three sequence analysis based methods. *Nucleic Acids Res* 34:1317–1325.
- Matthews BW. 1995. Studies on protein stability with T4 lysozyme. *Adv Protein Chem* 46:249–278.
- Miller MP, Kumar S. 2001. Understanding human disease mutations through the use of interspecific genetic variation. *Hum Mol Genet* 10:2319–2328.
- Minoshima S, Mitsuyama S, Ohtsubo M, Kawamura T, Ito S, Shibamoto S, Ito F, Shimizu N. 2001. The KMDb/MutationView: a mutation database for human disease genes. *Nucleic Acids Res* 29:327–328.
- Mooney SD, Klein TE. 2002. The functional importance of disease-associated mutation. *BMC Bioinformatics* 3:24.
- Moretti S, Armougom F, Wallace IM, Higgins DG, Jongeneel CV, Notredame C. 2007. The M-Coffee web server: a meta-method for computing multiple sequence alignments by combining alternative alignment methods. *Nucleic Acids Res* 35:W645–W648.
- Morris AL, MacArthur MW, Hutchinson EG, Thornton JM. 1992. Stereochemical quality of protein structure coordinates. *Proteins* 12:345–364.
- Ng PC, Henikoff S. 2001. Predicting deleterious amino acid substitutions. *Genome Res* 11:863–874.
- Ng PC, Henikoff S. 2003. SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Res* 31:3812–3814.
- Nuin PA, Wang Z, Tillier ER. 2006. The accuracy of several multiple sequence alignment programs for proteins. *BMC Bioinformatics* 7:471.
- Obradović Z, Peng K, Vucetic S, Radivojac P, Brown CJ, Dunker AK. 2003. Predicting intrinsic disorder from amino acid sequence. *Proteins* 53(Suppl 6):566–572.
- Obradović Z, Peng K, Vucetic S, Radivojac P, Dunker AK. 2005. Exploiting heterogeneous sequence properties improves prediction of protein disorder. *Proteins* 61(Suppl 7):176–182.
- Ohnishi S, Takano K. 2004. Amyloid fibrils from the viewpoint of protein folding. *Cell Mol Life Sci* 61:511–524.
- Pajunen M, Turakainen H, Poussu E, Peränen J, Vihinen M, Savilahti H. 2007. High-precision mapping of protein protein interfaces: an integrated genetic strategy combining en masse mutagenesis and DNA-level parallel analysis on a yeast two-hybrid platform. *Nucleic Acids Res* 35:e103.
- Parthiban V, Gromiha MM, Schomburg D. 2006. CUPSAT: prediction of protein stability upon point mutations. *Nucleic Acids Res* 34:W239–W242.
- Pei J, Kim BH, Tang M, Grishin NV. 2007. PROMALS web server for accurate multiple protein sequence alignments. *Nucleic Acids Res* 35:W649–W652.
- Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO. 1999. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc Natl Acad Sci USA* 96:4285–4288.
- Peng K, Vucetic S, Radivojac P, Brown CJ, Dunker AK, Obradović Z. 2005. Optimizing long intrinsic disorder predictors with protein evolutionary information. *J Bioinform Comput Biol* 3:35–60.

- Plempner RK, Wolf DH. 1999. Retrograde protein translocation: ERADication of secretory proteins in health and disease. *Trends Biochem Sci* 24:266–270.
- Poussu E, Vihinen M, Paulin L, Savilahti H. 2004. Probing the  $\alpha$ -complementing domain of *E. coli*  $\beta$ -galactosidase with use of an insertional pentapeptide mutagenesis strategy based on Mu in vitro DNA transposition. *Proteins* 54:681–692.
- Prilusky J, Felder CE, Zeev-Ben-Mordehai T, Rydberg EH, Man O, Beckmann JS, Silman I, Sussman JL. 2005. FoldIndex: a simple tool to predict whether a given protein sequence is intrinsically unfolded. *Bioinformatics* 21:3435–3438.
- Promponas VJ, Enright AJ, Tsoka S, Kreil DP, Leroy C, Hamodrakas S, Sander C, Ouzounis CA. 2000. CAST: an iterative algorithm for the complexity analysis of sequence tracts. Complexity analysis of sequence tracts. *Bioinformatics* 16:915–922.
- Raghava GP, Searle SM, Audley PC, Barber JD, Barton GJ. 2003. OXBench: a benchmark for evaluation of protein multiple sequence alignment accuracy. *BMC Bioinformatics* 4:47.
- Ramensky V, Bork P, Sunyaev S. 2002. Human non-synonymous SNPs: server and survey. *Nucleic Acids Res* 30:3894–3900.
- Rao VD, Misra S, Boronnikov IV, Anderson RA, Hurley JH. 1998. Structure of type IIa phosphatidylinositol phosphate kinase: a protein kinase fold flattened for interfacial phosphorylation. *Cell* 94:829–839.
- Reumers J, Maurer-Stroh S, Schymkowitz J, Rousseau F. 2006. SNPeff v.20: a new step in investigating the molecular phenotypic effects of human non-synonymous SNPs. *Bioinformatics* 22:2183–2185.
- Rocchia W, Sridharan S, Nicholls A, Alexov E, Chiabrera A, Honig B. 2002. Rapid grid-based construction of the molecular surface and the use of induced surface charge to calculate reaction field energies: applications to the molecular systems and geometric objects. *J Comput Chem* 23:128–137.
- Romero P, Obradović Z, Li X, Garner EC, Brown CJ, Dunker AK. 2001. Sequence complexity of disordered protein. *Proteins* 42:38–48.
- Rong SB, Väliaho J, Vihinen M. 2000. Structural basis of Bloom syndrome (BS) causing mutations in the BLM helicase domain. *Mol Med* 6:155–164.
- Rong SB, Vihinen M. 2000. Structural basis of Wiskott-Aldrich syndrome causing mutations in the WH1 domain. *J Mol Med* 78:530–537.
- Rousseau F, Schymkowitz J, Serrano L. 2006. Protein aggregation and amyloidosis: confusion of the kinds? *Curr Opin Struct Biol* 16:118–126.
- Sandberg WS, Schlunk PM, Zabin HB, Terwilliger TC. 1995. Relationship between in vivo activity and in vitro measures of function and stability of a protein. *Biochemistry* 34:11970–11978.
- Saqi MA, Sternberg MJ. 1994. Identification of sequence motifs from a set of proteins with related function. *Protein Eng* 7:165–171.
- Saunders CT, Baker D. 2002. Evaluation of structural and evolutionary contributions to deleterious mutation prediction. *J Mol Biol* 322:891–901.
- Schneider G, Fechner U. 2004. Advances in the prediction of protein targeting signals. *Proteomics* 4:1571–1580.
- Schymkowitz J, Borg J, Stricher F, Nys R, Rousseau F, Serrano L. 2005. The FoldX web server: an online force field. *Nucleic Acids Res* 33:W382–W388.
- Serrano L, Kellis Jr JT, Cann P, Matoschek A, Fersht AR. 1992. The folding of an enzyme. II. Substructure of barnase and the contribution of different interactions to protein stability. *J Mol Biol* 224:783–804.
- Sheinerman FB, Honig B. 2002. On the role of electrostatic interactions in the design of protein–protein interfaces. *J Mol Biol* 318:161–177.
- Shen B, Bai J, Vihinen M. 2008. Physicochemical feature-based classification of amino acid mutations. *Protein Eng Des Sel* 21:37–44.
- Shen B, Vihinen M. 2003. RankViaContact: ranking and visualization of amino acid contacts. *Bioinformatics* 19:2161–2162.
- Shen B, Vihinen M. 2004. Conservation and covariance in PH domain sequences: physicochemical profile and information theoretical analysis of XLA-causing mutations in the Btk PH domain. *Protein Eng Des Sel* 17:267–276.
- Shimizu K, Hirose S, Noguchi T. 2007a. POODLE-S: web application for predicting protein disorder by using physicochemical features and reduced amino acid set of a position-specific scoring matrix. *Bioinformatics* 23:2337–2338.
- Shimizu K, Muraoka Y, Hirose S, Tomii K, Noguchi T. 2007b. Predicting mostly disordered proteins by using structure-unknown protein data. *BMC Bioinformatics* 8:78.
- Shortle D, Stites WE, Meeker AK. 1990. Contributions of the large hydrophobic amino acids to the stability of staphylococcal nuclease. *Biochemistry* 29:8033–8041.
- Smith DK, Radivojac P, Obradović Z, Dunker AK, Zhu G. 2003. Improved amino acid flexibility parameters. *Protein Sci* 12:1060–1072.
- Sobolev V, Eyal E, Gerzon S, Potapov V, Babor M, Prilusky J, Edelman M. 2005. SPACE: a suite of tools for protein structure prediction and analysis based on complementarity and environment. *Nucleic Acids Res* 33:W39–W43.
- Sobolev V, Sorokina A, Prilusky J, Abola EE, Edelman M. 1999. Automated analysis of interatomic contacts in proteins. *Bioinformatics* 15:327–332.
- Stalker J, Gibbins B, Meidl P, Smith J, Spooner W, Hotz HR, Cox AV. 2004. The Ensembl Web site: mechanics of a genome browser. *Genome Res* 14:951–955.
- Stenson PD, Ball E, Howells K, Phillips A, Mort M, Cooper DN. 2008. Human Gene Mutation Database: towards a comprehensive central mutation database. *J Med Genet* 45:124–126.
- Stenson PD, Ball EV, Mort M, Phillips AD, Shiel JA, Thomas NS, Abeyasinghe S, Krawczak M, Cooper DN. 2003. Human Gene Mutation Database (HGMD): 2003 update. *Hum Mutat* 21:577–581.
- Stone EA, Sidow A. 2005. Physicochemical constraint violation by missense substitutions mediates impairment of protein function and disease severity. *Genome Res* 15:978–986.
- Strickler SS, Gribenko AV, Keiffer TR, Tomlinson J, Reihle T, Loladze VV, Makhataadze GI. 2006. Protein stability and surface electrostatics: a charged relationship. *Biochemistry* 45:2761–2766.
- Su CT, Chen CY, Hsu CM. 2007. iPDA: integrated protein disorder analyzer. *Nucleic Acids Res* 35:W465–W472.
- Suel GM, Lockless SW, Wall MA, Ranganathan R. 2003. Evolutionarily conserved networks of residues mediate allosteric communication in proteins. *Nat Struct Biol* 10:59–69.
- Sunyaev S, Lathe III W, Bork P. 2001a. Integration of genome data and protein structures: prediction of protein folds, protein interactions and “molecular phenotypes” of single nucleotide polymorphisms. *Curr Opin Struct Biol* 11:125–130.
- Sunyaev S, Ramensky V, Bork P. 2000. Towards a structural basis of human non-synonymous single nucleotide polymorphisms. *Trends Genet* 16:198–200.
- Sunyaev S, Ramensky V, Koch I, Lathe III W, Kondrashov AS, Bork P. 2001b. Prediction of deleterious human alleles. *Hum Mol Genet* 10:591–597.
- Tang H, Wyckoff GJ, Lu J, Wu CI. 2004. A universal evolutionary index for amino acid changes. *Mol Biol Evol* 21:1548–1556.
- Tartaglia GG, Cavalli A, Pellarin R, Caffisch A. 2005. Prediction of aggregation rate and aggregation-prone segments in polypeptide sequences. *Protein Sci* 14:2723–2734.
- Tavtigian S, Byrnes G, Goldgar D, Thomas A. 2008a. Classification of rare missense substitutions, using risk surfaces, with genetic and molecular epidemiology applications. *Hum Mutat* 29:1342–1364.
- Tavtigian S, Greenblatt M, Lesueur F, Byrnes G. 2008b. In silico analysis of missense substitutions using sequence alignment based methods. *Hum Mutat* 29:1327–1336.
- Terp BN, Cooper DN, Christensen IT, Jorgensen FS, Bross P, Gregersen N, Krawczak M. 2002. Assessing the relative importance of the biophysical properties of amino acid substitutions associated with human genetic disease. *Hum Mutat* 20:98–109.
- The International HapMap Consortium. 2003. The International HapMap Project. *Nature* 426:789–796.
- Thomas PD, Campbell MJ, Kejariwal A, Mi H, Karlak B, Daverman R, Diemer K, Muruganujan A, Narechania A. 2003. PANTHER: a library of protein families and subfamilies indexed by function. *Genome Res* 13:2129–2141.
- Thompson JD, Higgins DG, Gibson TJ. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* 22:4673–4680.
- Thompson JD, Plewniak F, Poch O. 1999. A comprehensive comparison of multiple sequence alignment programs. *Nucleic Acids Res* 27:2682–2690.
- Thompson MJ, Sievers SA, Karanicolas J, Ivanova MI, Baker D, Eisenberg D. 2006. The 3D profile method for identifying fibril-forming segments of proteins. *Proc Natl Acad Sci USA* 103:4074–4078.
- Thusberg J, Vihinen M. 2006. Bioinformatic analysis of protein structure–function relationships: case study of leukocyte elastase (ELA2) missense mutations. *Hum Mutat* 27:1230–1243.
- Thusberg J, Vihinen M. 2007. The structural basis of hyper IgM deficiency—CD40L mutations. *Protein Eng Des Sel* 20:133–141.
- Tomba P. 2002. Intrinsically unstructured proteins. *Trends Biochem Sci* 27:527–533.
- Tomba P, Csermely P. 2004. The role of structural disorder in the function of RNA and protein chaperones. *FASEB J* 18:1169–1175.
- Torkamani A, Schork NJ. 2007. Accurate prediction of deleterious protein kinase polymorphisms. *Bioinformatics* 23:2918–2925.
- Trovato A, Chiti F, Maritan A, Seno F. 2006. Insight into the structure of amyloid fibrils from the analysis of globular proteins. *PLoS Comput Biol* 2:e170.
- Trovato A, Seno F, Tosatto SC. 2007. The PASTA server for protein aggregation prediction. *Protein Eng Des Sel* 20:521–523.
- Ventura S, Zurdo J, Narayanan S, Parreño M, Mangues R, Reif B, Chiti F, Giannoni E, Dobson CM, Aviles FX, Serrano L. 2004. Short amino acid stretches can mediate amyloid formation in globular proteins: the Src homology 3 (SH3) case. *Proc Natl Acad Sci USA* 101:7258–7263.
- Vihinen M, Kwan SP, Lester T, Ochs HD, Resnick I, Väliaho J, Conley ME, Smith CIE. 1999. Mutations of the human BTK gene coding for bruton tyrosine kinase in X-linked agammaglobulinemia. *Hum Mutat* 13:280–285.

- Vihinen M, Nilsson L, Smith CIE. 1994a. Structural basis of SH2 domain mutations in X-linked agammaglobulinemia. *Biochem Biophys Res Commun* 205: 1270–1277.
- Vihinen M, Vetrie D, Maniar HS, Ochs HD, Zhu Q, Vořechovský I, Webster AD, Notarangelo LD, Nilsson L, Sowadski JM, Smith CIE. 1994b. Structural basis for chromosome X-linked agammaglobulinemia: a tyrosine kinase disease. *Proc Natl Acad Sci USA* 91:12803–12807.
- Vihinen M, Zvelebil MJ, Zhu Q, Brooimans RA, Ochs HD, Zegers BJ, Nilsson L, Waterfield MD, Smith CIE. 1995. Structural basis for pleckstrin homology domain mutations in X-linked agammaglobulinemia. *Biochemistry* 34:1475–1481.
- Vitkup D, Sander C, Church GM. 2003. The amino-acid mutational spectrum of human genetic disease. *Genome Biol* 4:R72.
- Vucetic S, Brown CJ, Dunker AK, Obradović Z. 2003. Flavors of protein disorder. *Proteins* 52:573–584.
- Vullo A, Bortolami O, Pollastri G, Tosatto SC. 2006. Spritz: a server for the prediction of intrinsically disordered regions in protein sequences using kernel machines. *Nucleic Acids Res* 34:W164–W168.
- Wagner M, Adamczak R, Porollo A, Meller J. 2005. Linear regression models for solvent accessibility prediction in proteins. *J Comput Biol* 12:355–369.
- Wallace IM, O'Sullivan O, Higgins DG, Notredame C. 2006. M-Coffee: combining multiple sequence alignment methods with T-Coffee. *Nucleic Acids Res* 34: 1692–1699.
- Wang Z, Moult J. 2001. SNPs, protein structure, and disease. *Hum Mutat* 17: 263–270.
- Ward JJ, Sodhi JS, McGuffin LJ, Buxton BF, Jones DT. 2004. Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *J Mol Biol* 337:635–645.
- Wheeler DL, Church DM, Federhen S, Lash AE, Madden TL, Pontius JU, Schuler GD, Schriml LM, Sequeira E, Tatusova TA, Wagner L. 2003. Database resources of the National Center for Biotechnology. *Nucleic Acids Res* 31:28–33.
- Williams AD, Shivaprasad S, Wetzel R. 2006. Alanine scanning mutagenesis of A $\beta$  (1–40) amyloid fibril stability. *J Mol Biol* 357:1283–1294.
- Wood LD, Parsons DW, Jones S, Lin J, Sjöblom T, Leary RJ, Shen D, Boca SM, Barber T, Ptak J, Silliman N, Szabo S, Dezso Z, Ustyanksky V, Nikolskaya T, Nikolsky Y, Karchin R, Wilson PA, Kaminker JS, Zhang Z, Croshaw R, Willis J, Dawson D, Shipitsin M, Willson JK, Sukumar S, Polyak K, Park BH, Pethiyagoda CL, Pant PV, Ballinger DG, Sparks AB, Hartigan J, Smith DR, Suh E, Papadopoulos N, Buckhaults P, Markowitz SD, Parmigiani G, Kinzler KW, Velculescu VE, Vogelstein B. 2007. The genomic landscapes of human breast and colorectal cancers. *Science* 318:1108–1113.
- Wootton JC. 1994. Non-globular domains in protein sequences: automated segmentation using complexity measures. *Comput Chem* 18:269–285.
- Word JM, Bateman Jr RC, Presley BK, Lovell SC, Richardson DC. 2000. Exploring steric constraints on protein mutations using MAGE/PROBE. *Protein Sci* 9:2251–2259.
- Word JM, Lovell SC, LaBean TH, Taylor HC, Zalis ME, Presley BK, Richardson JS, Richardson DC. 1999. Visualizing and quantifying molecular goodness-of-fit: small-probe contact dots with explicit hydrogen atoms. *J Mol Biol* 285:1711–1733.
- Worth CL, Bickerton GR, Schreyer A, Forman JR, Cheng TM, Lee S, Gong S, Burke DF, Blundell TL. 2007. A structural bioinformatics approach to the analysis of nonsynonymous single nucleotide polymorphisms (nsSNPs) and their relation to disease. *J Bioinform Comput Biol* 5:1297–1318.
- Yang ZR, Thomson R, McNeil P, Esnouf RM. 2005. RONN: the bio-basis function neural network technique applied to the detection of natively disordered regions in proteins. *Bioinformatics* 21:3369–3376.
- Yin S, Ding F, Dokholyan NV. 2007. Eris: an automated estimator of protein stability. *Nat Methods* 4:466–467.
- Yip YL, Famiglietti M, Gos A, Duek PD, David FP, Gateau A, Bairoch A. 2008. Annotating single amino acid polymorphisms in the UniProt/Swiss-Prot knowledgebase. *Hum Mutat* 29:361–366.
- Yoon S, Welsh WJ. 2004. Detecting hidden sequence propensity for amyloid fibril formation. *Protein Sci* 13:2149–2160.
- Yue P, Li Z, Moult J. 2005. Loss of protein structure stability as a major causative factor in monogenic disease. *J Mol Biol* 353:459–473.
- Yue P, Melamud E, Moult J. 2006. SNPs3D: candidate gene and SNP selection for association studies. *BMC Bioinformatics* 7:166.
- Zhou H, Zhou Y. 2002. Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. *Protein Sci* 11:2714–2726.