JANITA THUSBERG

Molecular Effects of Missense Mutations

Bioinformatics analysis of genetic defects

ACADEMIC DISSERTATION
To be presented, with the permission of
the Faculty of Medicine of the University of Tampere,
for public discussion in the Auditorium of Finn-Medi 1,
Biokatu 6, Tampere, on July 2nd, 2010, at 12 o'clock.

UNIVERSITY OF TAMPERE

UNIVERSITY
OF TAMPERE

ACADEMIC DISSERTATION
University of Tampere, Institute of Medical Technology
Tampere Graduate School in Biomedicine and Biotechnology (TGSBB)
Finland

Supervised by
Professor Mauno Vihinen
University of Tampere
Finland

Reviewed by
Professor Rita Casadio
University of Bologna
Italy
Docent Heikki Lehväslaiho
University of Helsinki
Finland

Distribution
Bookshop TAJU
P.O. Box 617
33014 University of Tampere
Finland

Tel. +358 40 190 9800
Fax +358 3 3551 7685
taju@uta.fi
www.uta.fi/taju
http://granum.uta.fi

Cover design by
Juha Siro

Acta Universitatis Tamperensis 1528
ISBN 978-951-44-8101-7 (print)
ISSN-L 1455-1616
ISSN 1455-1616

Acta Electronica Universitatis Tamperensis 968
ISBN 978-951-44-8102-4 (pdf)
ISSN 1456-954X
http://acta.uta.fi

Tampereen Yliopistopaino Oy – Juvenes Print
Tampere 2010

Contents

List of original communications

I. **Janita Thusberg** and Mauno Vihinen: Bioinformatic analysis of protein structure-function relationships: case study of leukocyte elastase (ELA2) missense mutations. Hum Mutat. (2006) 27: 1230-1243.

II. **Janita Thusberg** and Mauno Vihinen: The structural basis of hyper IgM deficiency – CD40L mutations. Protein Eng Des Sel. (2007) 20: 133-141.

III. Ilkka Lappalainen[*], **Janita Thusberg**[*], Bairong Shen and Mauno Vihinen: Genome wide analysis of pathogenic SH2 domain mutations. Proteins (2008) 72: 779-782.

IV. **Janita Thusberg** and Mauno Vihinen: Pathogenic or not? And if so, then how? Studying the effects of missense mutations using bioinformatics methods. Hum Mutat. (2009) 30: 703-714.

V. **Janita Thusberg**, Ayodeji Olatubosun and Mauno Vihinen: Performance of mutation pathogenicity prediction methods. Submitted.

[*] Joint first-authorship

Abbreviations

2D	two-dimensional
3D	three-dimensional
AA	amino acid
BCC	basal cell carcinoma
BTK	Bruton’s tyrosine kinase
CBD	chronic beryllium disease
CD40L	CD40 ligand
CMDB	central mutation database
dbSNP	The Single Nucleotide Polymorphism Database
DNA	deoxyribonucleic acid
GP	genotype-phenotype
HGMD	The Human Genome Mutation Database
HapMap	the Human Haplotype Map
HNE	human neutrophil elastase
Ig	immunoglobulin
JMML	juvenile myelomonocytic leukaemia
LSDB	locus-specific database
MCC	Matthews correlation coefficient
mRNA	messenger RNA
MSA	multiple sequence alignment
NK	natural killer
NMD	nonsense-mediated decay
NPV	negative predictive value
NS	Noonan syndrome
nsSNP	non-synonymous single nucleotide polymorphism
OMIM	Online Mendelian Inheritance in Man
PDB	Protein Data Bank
PTM	post-translational modification

PTP	protein tyrosine phosphatase
RNA	ribonucleic acid
SAS	solvent accessible surface
SCID	severe combined immunodeficiency
SH2	Src homology 2
SNP	single nucleotide polymorphism
TNF	tumor necrosis factor
UniProt	Universal Protein Resource
UPR	unfolded protein response
XHIGM	X-linked hyper-IgM syndrome
XLA	X-linked agammaglobulinemia
XLP	X-linked lymphoproliferative disease

Yhteenveto

Tietoa ihmisen geneettisistä polymorfioista kerääntyy nopeasti, mutta niiden mahdollisista yhteyksistä sairauksiin ja geneettisten sairauksien molekyylitason mekanismeista ei ymmärretä vielä riittävästi. Tämä johtuu siitä, että variaatiodataa kyetään tuottamaan nopeammin kuin analysoimaan. Kokeelliset tutkimusmenetelmät ovat usein työläitä ja aikaa vieviä. Laskennallisilla menetelmillä informaatiota voidaan tuottaa tehokkaammin, ja bioinformatiikan menetelmin tuotettua tietoa voidaan käyttää esimerkiksi kokeellisten tutkimusten suunnitteluun ja kiinnostavampien geenimuutosten priorisointiin.

Yhden emäksen muutokset eli pistemutaatiot DNA:ssa ovat yleisin ihmisten välisen geneettisen variaation muoto. Osa pistemutaatioista aiheuttaa aminohapon muuttumisen toiseksi proteiinissa, jonka aminohappokoostumuksen geeni määrittää. Noin puolet ihmisen perinnöllisistä sairauksista aiheutuu patogeenisista yhden aminohapon muutoksista. Yhden aminohapon muutokset voivat aiheuttaa moninaisia ja eriasteisia, ääritapauksessa vakavaan sairauteen johtavia rakenteen ja toiminnan muutoksia proteiineissa.

Tässä työssä tutkittiin yhden aminohapon muutosten laaja-alaisia vaikutuksia proteiineihin, ja analysoitiin siten perinnöllisten sairauksien molekyylitason syitä. Aminohappomuutosten bioinformatiikka-analyysille kehitettiin protokolla, jota sovellettiin ja kehitettiin sairauksien syitä tutkittaessa. Tutkimuksessa kehitetty protokolla toimii perustana jatkossa kehitettävälle uudelle tietokoneohjelmalle, joka on monipuolinen aminohappomuutosten vaikutusten analysointimenetelmä.

Useita laskennallisia menetelmiä aminohappomuutosten vaikutusten ennustamiseksi on jo kehitetty. Nämä ohjelmat perustuvat tutkittaviin proteiineihin liittyvän sekvenssi-informaation ja/tai proteiinirakenteiden analysointiin, ja ohjelmien tavoitteena on automatisoida yhden aminohapon muutosten vaikutusten tutkimista. Automatisointi olisi suureksi hyödyksi mutaatiotutkimukselle, jonka tavoitteena on selvittää, mitkä polymorfismit liittyvät geneettisiin sairauksiin. Tässä työssä tehtiin yhdeksän ennustusmenetelmän vertaileva tutkimus käyttämällä yli 60

000 neutraalin ja sairauden aiheuttavan aminohappomuutoksen aineistoa. Ohjelmien luotettavuuden havaittiin vaihtelevan merkittävästi, ja toimintaperiaatteeltaan samankaltaisilla ohjelmilla saatiin hyvin erilaisia tuloksia. Vertailun perusteella ohjelmat pystyttiin asettamaan paremmuusjärjestykseen, ja parhailla ohjelmilla voidaan saada riittävän luotettavia tuloksia aminohappomuutosten priorisoimiseksi jatkotutkimuksia varten. Tarkempien menetelmien kehittäminen on kuitenkin tarpeen, jotta ennestään tuntemattoman polymorfismin mahdollinen patogeenisuus voitaisiin luotettavasti ennustaa.

Tässä tutkimuksessa saavutettiin kiinnostavia näkökohtia muutamiin pistemutaatioiden aiheuttamiin perinnöllisiin sairauksiin. Näitä tuloksia voidaan hyödyntää kyseisten sairauksien molekyylitason syiden kokeellisessa tutkimuksessa. Mutaatioanalyysiprotokolla on perusteellinen menetelmä mutaatioiden vaikutusten tutkimiseksi, ja protokollan kehittäminen verkkopalveluksi mahdollistaa sen tehokkaan soveltamisen mutaatiotutkimuksessa. Uuden analysointiohjelman tavoitteena on tuottaa nykyisiä menetelmiä monipuolisempia ja laadukkaampia ennusteita pistemutaatioiden aiheuttamien aminohappomuutosten vaikutuksista.

Abstract

Available data on polymorphisms in the human genome are expanding rapidly, however knowledge of the possible disease association of polymorphisms and the molecular mechanisms of genetic disease is lagging due to the laborious and time-consuming nature of experimental studies. Bioinformatics studies can efficiently produce useful information to rationalise and guide further experimental study, and to shortlist the most interesting cases from the pool of accumulating data.

Some genetic variations, termed non-synonymous single nucleotide polymorphisms (nsSNPs), cause amino acid substitutios in the protein product of the gene. nsSNPs are the most common type of genetic variation among humans, and pathogenic nsSNPs, also termed missense mutations, account for approximately half of the allelic variants causative of hereditary disease. Amino acid substitutions may have diverse effects on protein structure and function, although some are functionally neutral.

In this study, the wide-ranging effects of amino acid substitutions were investigated at the protein level, and based on the analyses of missense mutations, the molecular basis of a number of hereditary diseases was elucidated. A protocol for the bioinformatics study of mutational effects was designed and implemented. The protocol serves as a basis for the development of a new service for predicting the effects of a missense mutation.

Several computational methods for predicting the possible pathogenicity of nsSNPs have been developed. These methods are based on evolutionary information and/or varying structural descriptors of the protein in question. These methods aim at automating the annotation process of nsSNP effects and therefore would be very useful for the mutation research community. In this study the performance of nine available prediction methods was evaluated using a dataset of over 60,000 missense mutations and polymorphisms. Significant differences in the prediction power of individual programs were observed, regardless of the apparent similarities between the programs. Some of the predictors perform well enough to be used in proritising

cases for further investigation; however more accurate methods are needed for reliable annotation of the putative effects of an nsSNP.

This study yielded interesting insights at the molecular level mechanisms of hereditary diseases, which can be utilised in further experimental studies. The protocol for mutation analysis is a comprehensive method for studying mutational effects and its development into a web service will provide the mutation research community a novel tool for efficient analysis beyond the scope of existing methods.

1. Introduction

Protein function is based on amino acid composition and properties governed by the detailed structure of the protein. Knowledge on protein structure-function relationships is therefore essential in finding the molecular basis for hereditary diseases and in predicting protein function from structure (and vice versa). Study of structure-function relationships is also used in medicine and pharmaceutical applications, where knowledge of protein function in health and disease is essential for understanding diseases and variations at the molecular level, and also in finding cures for various diseases.

Mutation design and experimental study of the changes they cause is laborious and time consuming. As a consequence, it is convenient and efficient to do the background work needed for mutation design and engineering of protein properties *in silico*. To engineer protein properties rationally and quickly, it is reasonable to model the mutations and probe the effects of changes on protein structure and function computationally, prior to the actual production and biochemical characterisation of engineered proteins. Because prediction of the properties of a mutated protein is based on many aspects of protein structure and function, it is rational to do it automatically using a computer program, with the capability to store and utilise masses of information needed for effective prediction.

Polymorphisms in the genome are responsible for phenotypic differences between humans and susceptibility to genetic disease. Through large scale efforts for identifying human genomic variations, such as the HapMap project (http://www.hapmap.org) (The International HapMap Consortium 2003), The Human Variome Project (http://www.humanvariomeproject.org) (Ring et al. 2006), The 1000 Genomes Project (www.1000genomes.org), The Cancer Genome Atlas (http://cancergenome.nih.gov), and whole genome association studies (Liu et al. 2006), available data on polymorphisms are accumulating rapidly in central databases such as The Single Nucleotide Polymorphism Database (dbSNP) (Sherry et al. 2001), The Human Genome Variation Genotype-to-Phenotype Database

(HGVbaseG2P) (Thorisson et al. 2009), the Swiss-Prot variant page (Yip et al. 2004), The Human Genome Mutation Database (HGMD) (Stenson et al. 2009), and many locus-specific databases (LSDBs) (Horaitis et al. 2007). However, because of the high-throughput nature of most of these efforts, many polymorphisms have not been experimentally characterised in terms of their possible disease association. Furthermore, the underlying mechanisms by which a genetic variant has a deleterious functional effect on its gene product and therefore causes disease are not yet fully understood. Because of the vast amount of variation data available, experimental study of each variant cannot be achieved in a reasonable timescale. Consequently, predictive analysis of the effects of polymorphisms on gene function is needed in order to prioritise the cases that require further study, to elucidate the molecular basis of hereditary diseases caused by missense mutations, and to gain a better understanding of the relationships between genetic and phenotypic variation, as well as protein structure and function. In this study, we examined the properties of missense mutations and mechanisms at the molecular level of their pathogenicity, in a number of human hereditary diseases. The applicability and reliability of bioinformatics methods for the analysis of mutations was studied and a procedure for the detailed study of missense variants is proposed.

2. Review of the literature

2.1 Genetic variation

2.1.1 Types of genetic variation

Genetic alterations are diverse and may have several kinds of effects at the phenotypic level. Point mutations, or single nucleotide polymorphisms (SNPs), are the most common type of genetic variation (The International HapMap Consortium 2003). Because only about 5% of the human genome codes for the production of proteins (The International Human Genome Sequencing Consortium 2001), most SNPs are found outside coding sequences. These variations may have effects in gene expression and regulation, by interrupting regulatory regions and affecting transcription factor binding. Coding SNPs, especially non-synonymous coding SNPs (nsSNPs, also referred to as missense mutations) are of particular interest, because as a result the amino acid sequence of the encoded protein is changed and thus a residue substitution may affect the structure and/or function of the protein.

Some nsSNPs, called nonsense mutations, cause truncation of the polypeptide by introducing a premature termination codon, which may lead to a drastic change in the length of the gene product. Furthermore, nonsense-mediated decay (NMD), where the absence of the gene product causes the decay of messenger RNA (mRNA), limits the synthesis of abnormal proteins (Chang et al., 2007; Holbrook et al. 2004; Thermann et al. 1998; Zhang et al. 1998). It has been estimated that about half of all nonsense mutations cause NMD (Han et al. 2007; Yamaguchi-Kabata et al. 2008). Loss of the termination codon caused by a mutation may lead to a similar process, referred to as non-stop decay, preventing translation (Frischmeyer et al. 2002; van Hoof et al. 2002).

Synonymous coding SNPs are nucleotide changes that due to the plasticity of the genetic code, do not lead to amino acid substitution. They can, however, affect the expression of the gene product by interfering with normal mRNA splicing, leading to abnormally short or long gene products. Synonymous coding SNPs may also cause alterations in mRNA folding and translation of the protein (Kudla et al. 2009).

Insertions and deletions in the coding regions of the genome cause variable changes in the length of the encoded polypeptide and thus may have major effects on the structure and/or function of the protein product. Other forms of genetic variation include gross insertions or duplications, and more rarely, gross deletions, complex rearrangements, and chromosomal aberrations, including rearrangements in genomic DNA and copy number variation.

2.1.2 nsSNPs and missense mutations

The human genome is estimated to have up to 200,000 non-synonymous coding variants (Cargill et al. 1999). nsSNPs occurring within the coding regions of genes lead to alterations in the amino acid sequence of their protein products, potentially causing changes in the structure and/or function of the protein. Certain coding variants, termed missense mutations, are known to cause highly penetrant, Mendelian-inherited pathological conditions. Missense mutations account for approximately half of all allelic variants underlying inherited human diseases (Hamosh et al. 2005; Krawczak et al. 2000; Stenson et al. 2003).

Unlike gross gene lesions, insertions, deletions, nonsense mutations or modified RNA splicing, which affect the length of a polypeptide, or determine whether a polypeptide is translated at all, missense mutations exert more subtle effects on protein structure and function. The consequences of missense mutations can be more difficult to predict because of their diverseness and because a single amino acid change may lead to multiple effects. The prediction of the pathogenicity of an nsSNP is based on the degree to which the function of the protein is impaired by the amino acid substitution, but it is further complicated by factors influencing the severity of the phenotype, such as the genetic background and the environment (Stone and Sidow 2005). Thousands of genes for rare, heritable Mendelian disorders have been identified, in which variation in a single gene is both necessary and sufficient to cause disease (Hamosh et al. 2005). Common disorders, in contrast, have proven much more challenging to study, as they are thought to be due to the combined effect of many different susceptibility genetic variants interacting with environmental factors.

In general, missense mutations that gain clinical attention usually change the physicochemical properties of the amino acid residue sufficiently to affect the function of the gene product (Krawczak et al. 1998; Stone and Sidow 2005), but the most severe mutations are likely to result in lethal phenotypes that cannot be inherited (Steward et al. 2003). At the same time, protein molecules are rather robust and can be quite tolerant to alterations in amino acid sequence (Pajunen et al. 2007; Poussu et al. 2004). In principle, an nsSNP can be deleterious either because it leads to disruption of a site that is directly involved in the function of a protein (e.g. a catalytic residue, a residue involved in ligand binding, or a residue that forms a critical interaction with another protein), or because it causes destabilisation of protein structure, leading to protein degradation, or the amino acid substitution abolishes protein function because of loss of the structural framework that enabled the functionality of the protein in the first place. In this study, the above-mentioned mutational types are referred to as functional mutations and structural mutations, respectively. In either case, pathogenic amino acid substitutions (missense mutations) tend to have special characteristics that distinguish them from those nsSNPs that cause no phenotypic effect (neutral variations). The prediction of the consequences of nsSNPs, in order to discriminate neutral variants from those causative of a disease phenotype, is a major research challenge as the rapid growth of genomic tools has produced vast amounts of information about genetic variation among individuals (Karchin 2009; Mooney 2005; Ng and Henikoff 2006; Steward et al. 2003).

2.2 Proteins

2.2.1 Protein structure

Proteins are linear polymers of amino acids and the distinct sequence of amino acid residues in a polypeptide chain determines the three-dimensional structure of the folded protein. The amino acid sequence is referred to as the primary structure.

Secondary structure is the local conformation of the polypeptide chain, which is stabilised by hydrogen bonding and the properties of peptide bonds between amino acid residues. The dominant secondary structures in proteins are α-helix and β-strand, which were predicted based on the known physical limitations of polypeptide chains prior to the experimental determination of protein structures (Pauling et al. 1951). These regular structures are interspersed with irregular, although generally ordered loops or coils. Disordered loop regions, referred to as random coils, do not achieve a stable structure. Loop regions are often located at the surface of the protein and in addition to simply serving as transitions between regular structures, they often harbour active sites in enzymes.

The global three-dimensional (3D) structure of a protein - the arrangement of the secondary structure elements in the polar solvent space - is referred to as the tertiary structure. The tertiary structure is locally governed by the interactions formed between amino acid residue side chains and globally most importantly by the hydrophobic effect (Tanford 1978), where residues with hydrophobic side chains are packed into the core of the protein, away from the solvent. In the hydrophobic core of the protein, the polarity of the polypeptide backbone is neutralised by hydrogen bonding in secondary structural elements. Buried polar residues form hydrogen bonds with other polar residues or the polypeptide backbone and in some cases with integral water molecules contained inside of the protein. Charged residues in the hydrophobic core form ionic interactions with residues of opposite charge. Disulphide bonds formed between cysteine residues are the only type of covalent interaction formed between amino acid residues. Disulphide bonds further stabilise tertiary structure, but they are not found in all proteins. The structure-stabilising role of interactions between buried residues has been known for long, but the contribution of surface residues to protein stability was thought to be negligible, until the rather recent observation that optimised surface charge-charge interactions have a stabilising effect on protein structure (Strickler et al. 2006).

Quaternary structure refers to the organisation of monomers in multisubunit proteins, stabilised by similar interactions as the secondary and tertiary structures. The interfaces between monomers are often thought to be hydrophobic and thus resemble the cores of globular proteins rather than typical surfaces. This has been shown to be true for homodimeric proteins, because they rarely occur as monomers, and hence their interaction surfaces are permanently buried within the protein-

protein complex. Heteromers, instead, often occur and function as monomers in solution and thus the interfaces of transient complexes exhibit a more hydrophilic quality (Janin and Chothia 1990; Janin et al. 1988; Jones and Thornton 1996). These interfaces could not be as hydrophobic as those of homodimers, because a large solvent exposed hydrophobic area on the protein would energetically be unfavourable (Jones and Thornton 1996).

2.2.2 Structural information in the analysis of mutations

Three-dimensional protein structure can provide additional information and evidence for a disease phenotype. Structural information is needed to fully understand the effects of missense mutations and the molecular disease mechanisms, or to rationalise the consequences of introduced mutations in protein engineering. The structural consequences of a very small proportion of the known mutations have been studied. Due to the difficulty in producing high quality 3D structures experimentally, a bioinformatics approach is very useful in predicting the structural effects of mutations. Mapping of an amino acid substitution into the known 3D structure can reveal whether the replacement is likely to have an impact on the normal folding or structural framework of the protein, e.g. when the substituting side chain is much larger than the original one and cannot be accommodated into the wild type structure, or whether the amino acid replacement destroys essential structure-maintaining contacts e.g. in the hydrophobic core of a protein, or has a destabilising impact on electrostatic interactions, interactions with ligands, or other features of a protein (Sunyaev et al. 2001; Wang and Moult 2001).

The number of proteins for which three-dimensional structures have been determined is rather small, thus limiting the extent to which structure-based prediction can be used. There are over 13,000 entries in genes with known sequence in Online Mendelian Inheritance in Man (OMIM) (Hamosh et al. 2005) (OMIM statistics page, February 2010), whereas the number of experimentally determined human protein structures in the Protein Data Bank (PDB) (Berman et al. 2000) is currently about 6,500 when redundancy is diminished so that similar proteins with 95% sequence identity are removed from the search. There are several sequence-based methods for the analysis of mutations that can be used when a 3D structure of the protein of interest is not available. Furthermore, protein structure prediction (2D and 3D) can provide valuable information when used as a basis for the study of structural effects of mutations (Baker and Sali 2001; Khan and Vihinen 2009; Saunders and Baker 2002).

2.3 The effects of missense mutations

2.3.1 Databases

Databases serve as the basis for mutation research. Mutation databases list mutations that are causative or highly penetrant for a particular inherited disorder. Central mutation databases (CMDBs), the most widely used being HGMD (Krawczak et al. 2000; Stenson et al. 2003) and OMIM (Hamosh et al. 2005), and COSMIC (Forbes et al. 2010) for somatic mutations in cancer, store information about genetic variations on the genomic scale and aim at collecting and curating mutations in all genes. Locus specific mutation databases (LSDBs), on the other hand, are specialised in certain genes or diseases. For the study of all mutations in a specific gene, it would be advisable to retrieve the mutations from the corresponding LSDB, if available, since LSDBs generally have more mutations compared to CMDBs (George et al. 2008). Conduit databases, such as HOWDY (Hirakawa 2002), MutDB (Mooney and Altman 2003), Phencode (Giardine et al. 2007), SAAPdb (Cavallo and Martin 2005) and KMDB/MutationView (Minoshima et al. 2001) which link information from CMDBs, LSDBs, genome browsers and protein databases have emerged as well, to integrate clinical and phenotypic information with genomic data and information about the gene product.

The Single Nucleotide Polymorphism Database dbSNP (Sherry et al. 2001), the Human Genome Variation Genotype-to-Phenotype Database (HGVbaseG2P), and the Human Haplotype Map (HapMap) (Frazer et al. 2007) aim to capture common allelic variants in the human population that usually have little or no functional consequence, that is, non-pathogenic polymorphisms. Protein databases such as the Universal Protein Resource (UniProt) (The Uniprot Consortium 2010) contain detailed information about gene products, partial lists of mutations and links to other data sources.

2.3.2 Functional sites

A missense mutation located at a site critical to protein function typically leads to a disease phenotype. A critical site may be a catalytic residue or a residue involved in ligand binding in an enzyme, or a residue involved in binding to partner molecules. The disease phenotype in these cases may arise because of loss or gain of function, or altered binding specificity or affinity in the protein, while the expression or stability of the protein product is not necessarily affected. For example, in glycogen storage disease type I, missense mutations affecting catalytic residues abolish glucose-6-phosphatase-α enzymatic function, but do not affect translation or stability of the protein (Chou and Mansfield 2008). Translation initiation codons can be affected by missense mutations as well, preventing the formation of the protein product, or translation may start at the next possible ribosome starting point, in which case the protein product would be truncated. The consequences of missense mutations affecting functional sites are rather straightforward to define when the protein in question is well known, because information regarding the critical residues is typically annotated in major protein databases, such as UniProt (The Uniprot Consortium 2010). The Catalytic Site Atlas is a specialised database for detailed annotations of known and predicted enzyme catalytic residues (Porter et al. 2004). For other proteins, critical functional sites can be predicted using multiple sequence alignments, as discussed in the following section. There are also programs available for the prediction of ligand or partner molecule binding sites in proteins, such as the FINDSITE method (Brylinski and Skolnick 2008), CASTp (Dundas et al. 2006), Q-Site Finder (Laurie and Jackson 2005), and Discern (Sankararaman and Sjolander 2008).

2.3.3 Sequence conservation

Disease-associated mutations change the function or the structural stability of a protein, whereas residue differences between evolutionarily related proteins usually conserve protein structure and function (Steward et al. 2003; Vitkup et al. 2003). Pathogenic mutations tend to occur at positions conserved between species in evolution (Ferrer-Costa et al. 2002; Miller and Kumar 2001; Mooney and Klein 2002; Shen and Vihinen 2004; Steward et al. 2003; Sunyaev et al. 2000), and classically, highly conserved positions in multiple sequence alignments often point to functional sites (Capra and Singh 2007; Casari et al. 1995; Chung et al. 2006; Hu et al. 2000; Lichtarge et al. 1996; Panchenko et al. 2004; Zhou and Shan 2001; Zvelebil et al. 1987). When considering structural mutations, the level of conservation of the physicochemical properties between the wild type and the substituting amino acid has an effect on the pathogenicity of the mutation, so that conservative substitutions tend to be less frequently pathogenic than those significantly altering the residue properties such as charge, hydropathy, or size (Briscoe et al. 2004; Khan and Vihinen 2007; Miller and Kumar 2001; Stone and Sidow 2005; Tang et al. 2004). The hydrophobic nature of residues located in core of a protein especially tends to be conserved, and these residues can usually be identified in multiple sequence alignment.

On the contrary, there is an under-abundance of disease-causing mutations occurring at positions that change in evolution (Briscoe et al. 2004; Miller and Kumar 2001). Consequently, sequence conservation and phylogenetic studies are powerful for the prediction of functionally and structurally important residues in proteins. However, mouse genome data reveal many disease-associated mutations in humans that are wild-type residues in mouse orthologues (Waterston et al. 2002), so it cannot be directly assumed that because the same residue type appears at equivalent positions in close homologues, it will not lead to disease in humans. This phenomenon is partially explained by the fact that co-evolution of the sites vital to the structure and/or function of a protein is rather common. When a critical site is mutated, a compensating mutation occurs at a site that is functionally, energetically of physically linked to that position. Analysis of covariant positions in multiple sequence alignments may thus reveal conserved positions that are relevant to the function or structure of a protein.

2.3.4 Signal peptides

Signal peptides are sequences that mediate the targeting and translocation of proteins to the endoplasmic reticulum for subsequent processing. Amino acid substitutions in signal peptides can result in impaired protein targeting and thereby alter or abolish protein function. Mutations in signal peptides are rare (Laurila and Vihinen 2009), but several inherited diseases are known to be caused by mutations in signal peptides. These include among others, familial isolated hypoparathyroidism (Arnold et al. 1990; Karaplis et al. 1995), thyroxine-binding globulin deficiency (Fingerhut et al. 2004), Crigler-Najjar type II (Seppen et al. 1996), coagulation factor X deficiency (Racchi et al. 1993), familial central diabetes inspidus (Ito et al. 1993) and familial hypocalciuric hypercalcemia (Pidasheva et al. 2005). Methods have been developed for predicting signal peptides (reviewed in (Schneider and Fechner 2004)) and for discriminating between deleterious and neutral signal peptide variants (Hon et al. 2009; Laurila and Vihinen 2009).

2.3.5 Post-translational modification sites

Protein post-translational modifications (PTMs) such as phosphorylation, glycosylation, methylation, acetylation, lipid modifications and ubiquitylation, have wide-ranging effects on protein function and interactions with other molecules, and are thereby central to cellular behaviour and responses. In some cases, a residue side chain modification is essential for the proper formation of the protein tertiary structure. More often, PTMs have no effect on protein fold, but have a role in protein function or localisation within the cell, on the cell membrane or in the extracellular matrix. A missense mutation may abolish a PTM site by introducing an amino acid that cannot be modified, or altering the neighbouring residues so that the PTM site cannot be recognised, thereby leading to abnormal protein function. A missense mutation-induced gain of PTM is another possible disease mechanism, leading to protein destabilisation, changes in protein interactions, catalytic properties or other protein functions. Mutational defects in post-translational modification have been found in various disease phenotypes, e.g. cancer (Benzeno et al. 2006; Bode and Dong 2004; Lim 2005; Radivojac et al. 2008), primary immunodeficiencies (Vogt et al. 2005; Vogt et al. 2007), haemophilia (Aly et al. 1992), cystic fibrosis (Hämmerle et al. 2000) and many more.

2.3.6 Solvent accessibility

Buried positions are more sensitive to pathogenic mutations than positions on the surface of the protein, because alterations in such positions, especially in the hydrophobic core of a protein, have the potential to cause greater disruption of the overall structure of a protein (Steward et al. 2003; Sunyaev et al. 2000; Terp et al. 2002; Vitkup et al. 2003). Residues at these positions form critical stability-maintaining contacts with other residues and these may be disrupted when a residue is altered. Changes in the size (Buckle et al. 1996; Eriksson et al. 1992; Liu et al. 2000; Loladze et al. 2002; Otzen et al. 1995; Shortle et al. 1990), hydrophobicity (Liu et al. 2000; Matthews 1993; Shortle et al. 1990), or charge of the residue side chain at buried positions usually have an effect on the structural stability of the protein (Chasman and Adams 2001; Sunyaev et al. 2001). Mutations at solvent accessible sites might interfere with the interactions a protein forms with other molecules, or they may contribute to the solubility or stability of a protein (Gribenko et al. 2009; Grimsley et al. 1999; Strickler et al. 2006; Sunyaev et al. 2001; Wang and Moult 2001).

2.3.7 Interactions between residues and conformational stability

In order to perform its biological tasks, a protein must typically fold to its characteristic globular conformation. The conformational stability of a protein is defined as the free energy change, ΔG, for the reaction folded ↔ unfolded, under physiological conditions. The folded conformations are in general only 5 to 10 kcal/mol more stable than the biologically inactive, unfolded conformations (Pace 1990). Several forces contribute to the small net conformational stability, conformational entropy being the main destabilising force. The most important stabilising forces are hydrogen bonding and the hydrophobic effect. Folding of proteins is governed by the burial of side chains in the interior of the molecule, out of contact with water, and the formation of intramolecular contacts between side chains. The main secondary structural elements, α-helices and β-strands, are formed by hydrogen bonding among main chain polar groups, and hydrogen bonds among side chains contribute to the stability of the tertiary structure of the protein (Eswar and Ramakrishnan 2000). A missense mutation that alters the physicochemical properties of an amino acid residue will naturally have an effect on the contacts formed between residues in a protein and may thereby cause alterations in folding. A frequent consequence of missense mutations is that the mutant protein is correctly folded but less stable, or the mutant protein may be in a stable conformation with a structure slightly different from the native protein, in some cases different enough to cause the protein to be dysfunctional (Thomas et al. 1995). It has been estimated that over 80 % of mutations that affect protein function, do so through disruption of protein stability (Wang and Moult 2001).

The types of interactions that may be disrupted as a consequence of an amino acid substitution include hydrogen bonds (Shirley et al. 1992), hydrophobic (Eriksson et al. 1992; Matthews 1995) and van der Waals interactions (Eriksson et al. 1992; Xu et al. 1998), disulfide bonds (Betz 1993) and electrostatic interactions (Horovitz et al. 1990). Examples of missense mutations causing alterations in aforementioned interactions and thereby loss of molecular function are shown in Figure 1.

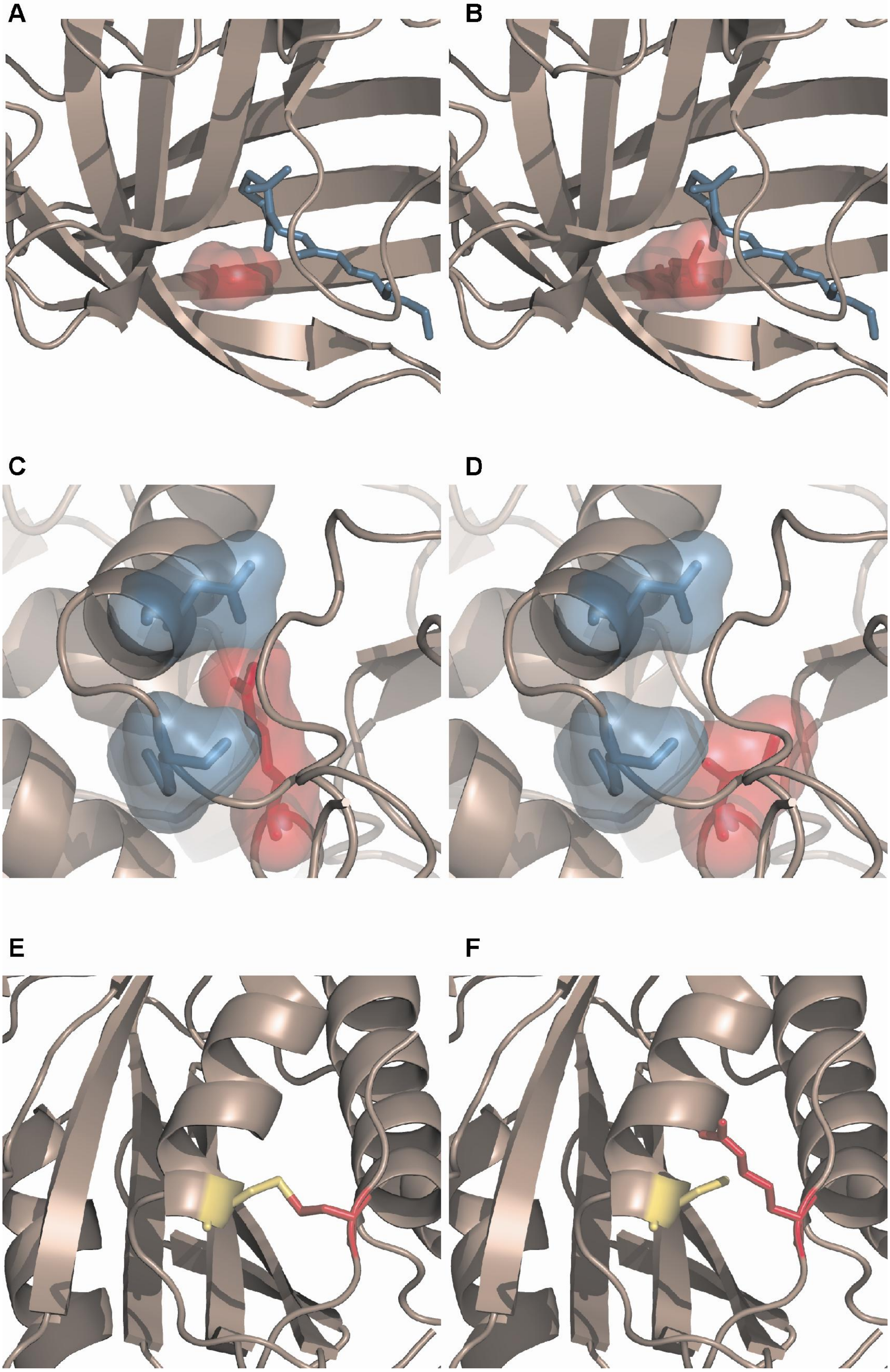

A
B
C
D
E
F

Figure 1. Examples of the ways in which structural effects of amino acid substitutions cause disease. A-B: Hindrance of ligand binding. Retinol (turquoise) binds in the centre of retinol-binding protein 4 (PDB ID 1BRP). A missense mutation affecting glycine 75 (magenta) causes vitamin A deficiency (Seeliger et al. 1999). The substitution of G75 by aspartic acid, as a result of the mutation, interferes with ligand binding by causing steric clashes between D75 and the retinol molecule. C-D: Loss of electrostatic interactions and hydrogen bonds. A missense mutation affecting R252 (magenta) in coagulation factor XIII (PDB ID 1GGT) leads to congenital factor XIII deficiency. R252 forms a salt bridge with D243 (turquoise, upper residue), and hydrogen bonds with the main chain carbonyl of M247 (turquoise, lower residue). The interactions are lost due to the substitution of R252 by isoleucine (magenta), leading to an unstable structure (Mikkola et al. 1996). E-F: Breakage of disulphide bond. The cysteines 509 (magenta) and 695 (yellow) form a covalent disulphide bond in wild type von Willebrand factor domain A1 (PDB ID 1SQ0). The substitution C509R causing type IIA von Willebrand disease leads to the breakage of the bond, causing the protein to be inactive (Lavergne et al. 1992).

2.3.8 Substitutions involving glycine and proline residues

The mutations affecting or introducing glycine and proline residues can be thought of as a distinct group of mutations because of the special characteristics of these amino acid residues. Glycine plays a very important role structurally because with only a hydrogen atom as a side chain, it can adopt a much larger range of conformations than other residues, providing structural flexibility that is lost upon mutation. Due to the wide range of possible backbone dihedral angles glycines are found in reverse turns (Rose et al., 1985, and substitution of a glycine residue forming this type of turn by any other residue is expected to destabilise the protein or cause the protein to adopt a different fold, at least locally (Pakula and Sauer 1989). Owing to the small size of the side chain, substitutions affecting buried positions normally occupied by glycine often cannot be fitted into the structure without rearrangements of the structure (Liu et al. 2000). On the other hand, mutations that introduce glycines may create cavities inside hydrophobic parts of the protein, causing destabilisation (Eriksson et al. 1992; Matthews 1995).

The cyclic structure of the proline side chain locks its φ backbone dihedral angle, causing it to be exceptionally rigid conformationally. Introduction of a proline residue causes altered strain in the polypeptide backbone and introduction of prolines at positions that require main chain torsional angles significantly different from those characteristic of proline has a destabilising effect (MacArthur and Thornton 1991; Pakula and Sauer 1989). Prolines are commonly known to disrupt secondary structure because in α-helices and β-strands the introduction of a pyrrolidine ring often causes steric clashes to neighbouring residue side chains (Schimmel and Flory 1968). In addition, as proline lacks a peptide-NH group, it is incapable of forming main chain hydrogen bonds important for 2D structure formation and stabilisation. Prolines are commonly found in turns because of the bend they cause in the polypeptide backbone and amino acid substitutions affecting these prolines may lead to structural rearrangements.

2.3.9 Structural disorder

Many globular proteins contain disordered segments and some proteins are intrinsically disordered. Mutations can, however, introduce disorder into ordered structures, thereby affecting protein function. For example, amino acid substitutions in the heat shock protein HSP22 have been shown to increase unordered structure and decrease the chaperone-like activity of the protein (Kasakov et al. 2007). On the other hand, a mutation-induced increase in ordered structure at intrinsically disordered parts of a protein might also affect the functionality of a protein. Disordered segments are important in for example, molecular recognition and interactions (Mészáros et al. 2007). Many eukaryotic regulatory proteins are intrinsically disordered and acquire a folded structure upon binding to their target molecule (Wright and Dyson 1999). Several prediction methods for structural disorder in proteins have been developed, and disorder, as well as the prediction of disorder, is further discussed in recent review articles (Bourhis et al. 2007; Dosztányi et al. 2007; Dunker et al. 2008; Uversky et al. 2008).

2.3.10 Misfolding and aggregation

Research of protein misfolding and aggregation is progressively gaining increasing attention, largely because of its impact on the understanding of the molecular mechanisms underlying widespread pathologies involving amyloid formation such as Alzheimer's (Selkoe 1996) or Parkinson's (Trojanowski and Lee 1998) disease. Aggregation is a common characteristic of polypeptide chains, involving the irreversible interaction of two or more denatured protein molecules leading to precipitation of protein. Missense mutations that trigger protein aggregation have been shown to be associated with an increasing number of pathologies (Bucciantini et al. 2004; Chiti et al. 2003; Chiti et al. 1999; Fandrich et al. 2001; Guijarro et al. 1998; Harris and True 2006; Keage et al. 2009; Khemtemourian et al. 2008; Robinson 2008; Yankner and Lu 2009). The formation of aggregates is triggered by the destabilisation and opening of the native protein structure, which exposes aggregation-prone regions previously buried inside the structure. The amyloidogenic sequence stretches can then nucleate the aggregation reaction (Dobson 2004; Ventura et al. 2004). The composition and primary structure of a protein determine to a large extent its propensity to aggregate, which is why even small alterations such as missense mutations may have a considerable effect in the solubility and aggregation propensity of a protein (Esteras-Chopo et al. 2005). Computational prediction of mutation-induced changes in aggregation has been shown to be successful in recent rational mutagenesis studies where aggregation propensities of proteins were altered (Cerdà-Costa et al. 2007; Fowler et al. 2005; Luheshi et al. 2007).

2.3.11 Electrostatic surface potentials

The local and global electrostatic surface potentials are essential for protein structure, function and binding to partner molecules. The surface electrostatics of proteins are the thermodynamic and kinetic rate-increasing steering force for interactions between proteins and other molecules. Ion pairing and other favourable electrostatic interactions also influence the structure of complexes. Patches of electrostatic potential on proteins are often indicators of a binding surface, typically to a molecule with a potential of opposite sign (Honig and Nicholls 1995). Protein-

protein interfaces often contain binding “hot spots” that contribute mostly to binding free energy. The hot spots consist of structurally conserved charged and polar residues surrounded by water-occluding hydrophobic residues (Bogan and Thorn 1998; Clackson and Wells 1995; Hu et al. 2000; Ma et al. 2003). In addition to intermolecular complex formation, electrostatic interactions can be formed between domains within a single protein as well. Not all molecular interfaces are charged, however. Some interfaces, especially homodimer binding surfaces, are dominated by hydrophobic residues (Jones and Thornton 1996).

Missense mutations that affect protein surface electrostatics may thus have diverse effects, ranging from changes in folding or stability, to alterations in partner/ligand binding affinity and specificity, and thus function of the protein. For example, allelic variants in the *HLA-DPB1* gene, known to increase susceptibility to chronic beryllium disease (CBD), were analysed by studying electrostatic surface potentials (Snyder et al. 2003). The probability of developing CBD estimated from epidemiological studies was found to correlate with the degree of change in the surface charge of the human leukocyte antigen binding pocket, predicted to alter the innate specificity of binding, thus elucidating genotype-phenotype correlations for the variants and mechanism for the disease. In another study, effects of cancer associated mutations on surface electrostatics and protein-protein interactions of the tumor suppressor p53 regulator MDM2 were predicted computationally (Lee et al. 2007) and the predictions verified experimentally (Brown et al. 2008). These studies indicated that the differential binding of phosphorylated and unphosphorylated p53 by its negative regulator MDM2 is based on charge repulsion between the two molecules induced by p53 phosphorylation. Charge-altering mutations of MDM2 lead to enhanced binding to phosphorylated p53, preventing the normal upregulation of p53 upon cellular stress, DNA damage and hypoxia.

2.3.12 Predictors of pathogenicity

Many efforts have been made to develop algorithms for the prediction of the functional impact of amino acid substituting variants that have not been characterised experimentally. These methods are based on calculations of different combinations of the above-mentioned properties of amino acids, amino acid sequences and protein structures, sometimes complemented with annotations from protein or mutation databases, such as SwissProt or OMIM. However, a method taking all aspects of the possible mechanisms of pathogenicity into account does not yet exist. The programs share a similar logic in classification (Figure 2), but are based on different machine learning techniques such as neural networks, random forests or support vector machines (Bao and Cui 2005; Bromberg and Rost 2007; Calabrese et al. 2009; Capriotti et al. 2006; Ferrer-Costa et al. 2005; Li et al. 2009), rule-based classification (Ramensky et al. 2002), or mathematical operations (Ng and Henikoff 2001), and the attributes used for describing the amino acid substitution vary from program to program.

Some methods for the prediction of missense mutation pathogenicity are based solely on sequence-level information and multiple sequence alignments (MSAs) (Ng and Henikoff 2001; Thomas et al. 2003). The advantage of these methods is the possibility to include those proteins in the analysis that lack a defined 3D structure. One would intuitively assume that the addition of structural parameters would improve the prediction, but this has not always been the case. Sequence conservation-based prediction methods have been shown to perform equally well with those methods that include structural parameters, when there is a sufficient number of homologous sequences available (Bao and Cui 2005; Saunders and Baker 2002). According to other studies (Bromberg and Rost 2007; Krishnan and Westhead 2003), inclusion of structural descriptors in the prediction improves predictor performance. However, the sequence-based approach for the study of pathogenicity of missense mutations has the disadvantage that it provides no direct insight into the underlying molecular mechanism of disease, although being powerful in distinguishing pathogenic mutations from benign variants.

Wang and Moult developed a structure-based model to evaluate the effect of amino acid substitutions based on a set of rules that take into account hydrophobic burial, backbone strain, over packing and electrostatic interactions (Wang and Moult 2001; Yue et al. 2005). Sunyaev and others predicted the effects of missense mutations using another set of rules, comprising functional information, hydrophobic propensity, side chain volume change and transmembrane location, together with sequence and phylogenetic information (Ramensky et al. 2002; Sunyaev et al. 2001). Yet another contemporary method developed by Chasman and

Adams utilised a set of similar structural parameters combined with phylogenetic information (Chasman and Adams 2001). A plethora of methods for the prediction of the pathogenicity of missense mutations have been developed since these pioneering studies.

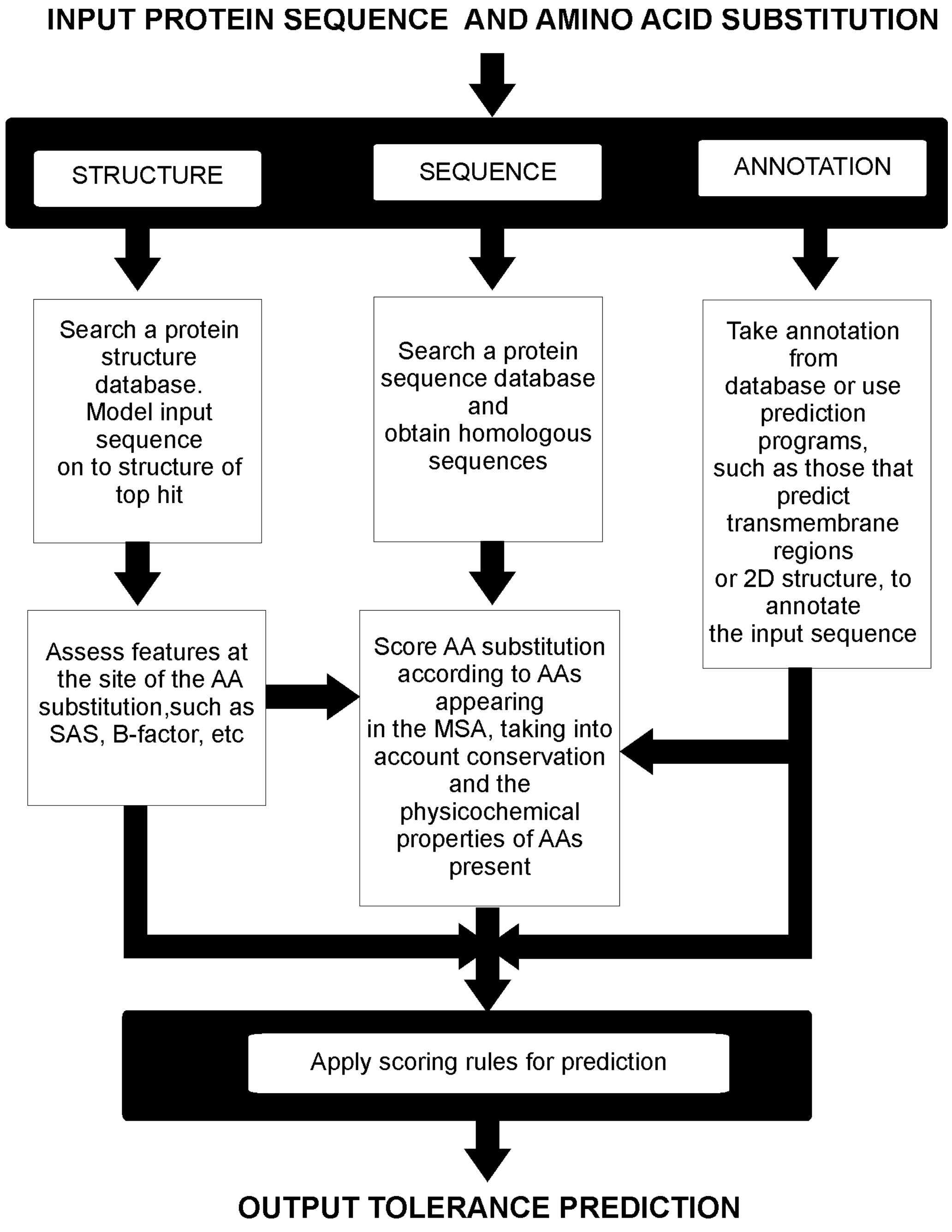

INPUT PROTEIN SEQUENCE AND AMINO ACID SUBSTITUTION
STRUCTURE
SEQUENCE
ANNOTATION
Search a protein structure database. Model input sequence on to structure of top hit
Search a protein sequence database and obtain homologous sequences
Take annotation from database or use prediction programs, such as those that predict transmembrane regions or 2D structure, to annotate the input sequence
Assess features at the site of the AA substitution,such as SAS, B-factor, etc
Score AA substitution according to AAs appearing in the MSA, taking into account conservation and the physicochemical properties of AAs present
Apply scoring rules for prediction
OUTPUT TOLERANCE PREDICTION

Figure 2. General function of prediction methods for missense variant pathogenicity. All the methods take protein sequence and amino acid substitution to be analysed as input, and give a tolerance prediction as output, after classifying the variant based on different procedures. Some programs operate solely based on MSAs (middle path) and others include structural information (left-hand path) and/or functional annotations (right-hand path) in the prediction.

2.3.13 Previous studies of the effects of missense mutations

Prediction of the molecular effects of disease-causing missense mutations by bioinformatics methods has been implemented in numerous recent studies. Many groups have analysed the pathogenicity of variants in the breast cancer susceptibility genes *BRCA1* and *BRCA2* by bioinformatics methods, or by using a combination of experimental and computational methods (Abkevich et al. 2004; Carvalho et al. 2009; Carvalho et al. 2007; Fleming et al. 2003; Goldgar et al. 2004; Karchin et al. 2007; Mirkovic et al. 2004; Rajasekaran et al. 2008; Rajasekaran et al. 2007; Williams et al. 2003; Williams and Glover 2003). Computational methods have been used to assess the pathogenicity of genetic variants in for example the *VHL* (Rajasekaran et al. 2008) and *ABL1* genes (George Priya Doss et al. 2008), and to evaluate the mechanism of pathogenicity of missense mutations in Type 2 diabetes mellitus (Sharma et al. 2005) and galactosemia (Facchiano and Marabotti 2010). Bioinformatics analysis has also been applied to study all mutations related to a specific group of proteins. Savas and coworkers predicted the pathogenicity of variants in 77 cell cycle proteins and their interaction partners (Savas et al. 2005), and Shen and others analysed variants in 45 cytokine proteins (Shen et al. 2006).

Understanding the molecular consequences of the mutations that cause human genetic disease remains an important research challenge (Karchin 2009; Mooney 2005; Ng and Henikoff 2006; Steward et al. 2003). Until now, the research has concentrated mainly on using just one or a few methods per study, but our motivation has been to attain more reliable results by utilising a more extensive set of prediction methods in the analysis of mutations. Our view is that the problem of elucidating the molecular level mechanisms of missense mutation pathogenicity should be approached from multiple perspectives, as the effects of amino acid substitutions on protein structure and function are diverse.

2.4 Genetic diseases analysed in this study

The molecular basis of a number of hereditary disorders was studied by the bioinformatics approach developed herein. The features of the diseases are summarised in Table 1 (on page 46).

2.4.1 Cyclic and congenital neutropenia

Mutations in the neutrophil elastase gene (*ELA2*) are causative of both cyclic and congenital forms of neutropenia, an autosomal dominant immunodeficiency, characterised by oscillating or decreased levels of neutrophils in the blood (Dale et al. 2000; Horwitz et al. 1999). The mutations lead to dysfunctionality of the protein product human neutrophil elastase (HNE), which is a serine protease expressed in azurophil granules in neutrophils and also extracellular space at sites of inflammation. HNE is responsible, among other proteases in the azurophil granules, for the degradation of objects internalised by phagocytosis (Boxer and Morganroth 1987; Lehrer et al. 1988) and for the degradation of various proteins in the extracellular space (Bach-Gansmo et al. 1996; Gillis et al. 1997; Weiss 1989; Wintroub et al. 1980). Neutropenia patients suffer from opportunistic and sometimes life-threatening infections, due to compromised immune system function.

2.4.2 X-linked hyper-IgM syndrome (XHIGM)

XHIGM is a primary immunodeficiency caused by mutations in the gene encoding CD40 ligand (CD40L), a protein expressed on T cell membranes. The interaction of CD40L with its receptor CD40, expressed on B cells, is essential for lymphocyte signalling, leading to B cell maturation and antibody isotype class switching (Kroczek et al. 1994). The mutations in CD40L result in an inability of the protein to bind to its receptor and consequently, interfere with the signalling cascade leading to the activation of several genes involved in B cell proliferation and antibody production (Allen et al. 1993). XHIGM is characterised by low levels or the absence of IgG, IgA and IgE, and normal or elevated IgM in the serum, causing the patients to be highly susceptible to recurrent bacterial infections and prone to autoimmune diseases and neutropenia (Fuleihan et al. 1993; Levy et al. 1997; Notarangelo et al. 1992).

2.4.3 Diseases caused by mutations in Src homology 2 (SH2) domains

SH2 domains are usually found in multidomain proteins, involved in cellular signalling pathways. These domains bind to their specific phosphopeptide targets and thus function in molecular assembly of activated complexes. Particular SH2 domains form intramolecular interactions, which regulate enzyme activity (Machida and Mayer 2005; Schlessinger and Lemmon 2003). Missense mutations in the genes that encode SH2 domain-containing proteins cause various diseases, owing to the versatile roles of SH2 domains in cellular signalling.

2.4.3.1 X-linked agammaglobulinemia (XLA)

XLA is caused by mutations in the gene encoding Bruton’s tyrosine kinase (BTK), a tyrosine kinase participating in multiple signalling pathways involved in B lymphocyte maturation (Lindvall et al. 2005). The disease is characterised by failure to produce mature B cells and associated with a failure of immunoglobulin heavy chain rearrangement, which leads to recurring bacterial infections in patients (Bruton 1952; Rawlings and Witte 1994).

2.4.3.2 X-linked lymphoproliferative syndrome (XLP)

The gene defective in XLP encodes a protein SH2D1A which is expressed mainly in T cells and natural killer (NK) cells (Coffey et al. 1998; Nichols et al. 1998). The protein functions as a component of a signalling cascade that leads to activation of T cells upon their association with antigen presenting cells. SH2D1A affects downstream signalling in several ways and therefore its dysfunction leads to the broad clinical spectrum of XLP (Sayos et al. 1998). XLP involves extreme sensitivity to infection with Epstein-Barr virus, which results in a complex phenotype characterised by mononucleosis, alterations in concentrations of serum immunoglobulins and malignant lymphoma (Purtilo 1981).

2.4.3.3 Severe combined immunodeficiency (SCID)

Mutations in the gene encoding ZAP-70, a tyrosine kinase involved in T cell receptor signalling (Chan et al. 1992), lead to a rare, autosomal recessive form of SCID (Chan et al. 1994). The ZAP-70 deficiency is characterised by the selective absence of CD8$^+$ T cells and an abundance of CD4$^+$ T cells that are unresponsive to T cell receptor mediated stimuli (Arpaia et al. 1994). Patients suffer from chronic diarrhoea, persistent candidiasis and severe pulmonary infections.

2.4.3.4 Noonan syndrome (NS)

NS (Noonan 1968) is caused by mutations in the gene *PTPN11*, which encodes the protein SHP-2 (Allanson 1987). SHP-2 is a tyrosine phosphatase that acts as a component in several signalling pathways involved in the control of developmental processes, haematopoiesis and metabolism (Qu et al. 1997; Qu et al. 1998; Saxton et al. 2000; Saxton et al. 1997; Tang et al. 1995; Zhang et al. 2004). NS is an autosomal dominant syndrome characterised by congenital heart defects, short stature, thrombocytopenia and a characteristic configuration of facial features (Kitchens and Alexander 1983; Lemire 2002; Limal et al. 2006).

2.4.3.5 Juvenile myelomonocytic leukaemia (JMML)

Mutations in the *PTPN11* gene, which is affected in Noonan syndrome have been shown to cause JMML (Bentires-Alj et al. 2004; Tartaglia et al. 2003). JMML is a paediatric myelodysplastic syndrome characterised by excessive proliferation of myelomonocytic cells (Hasle et al. 1999).

2.4.3.6 Severe insulin resistance

A missense mutation leading to an amino acid substitution in the SH2 domain of PI3-kinase, a signal transduction protein linking insulin to many of its cellular responses (Cohen 2006), has been found to cause severe insulin resistance (Baynes et al. 2000).

2.4.3.7 Basal cell carcinoma (BCC)

BCC is the most frequent skin cancer in the white population, usually occurring sporadically, but a subset of cases have been shown to be linked to genetic disorders (Bodak et al. 1999; Goeteyn et al. 1994; Gorlin 1987). Mutations in the gene encoding the GTPase-activating protein, RasGAP, have been found in a subset of BCC cases (Friedman et al. 1993).

2.4.3.8 STAT1 deficiency

The STAT1 protein mediates interferon signalling as part of the JAK/STAT1 pathway (Ramana et al. 2000). The complete STAT1 deficiency caused by mutations in the *STAT1* gene involves impaired response to interferon γ, leading to severe viral disease and mycobacteriosis (Dupuis et al. 2003).

2.4.3.9 Growth hormone insensitivity with immunodeficiency

Defects in the STAT5B SH2 domain lead to growth hormone insensitivity with immunodeficiency (Kofoed et al. 2003). STAT5B is a component of the growth hormone signalling pathway that leads to stimulation of insulin-like growth factor I gene transcription (Woelfle et al. 2003).

3. Aims of the study

The aims of this study were to:

1) Broaden our knowledge on protein structure-function relationships and genotype-phenotype correlations in human hereditary diseases, by studying known mutations and elaborating their effects on protein properties in silico (I-III).
2) Design and implement a protocol for the bioinformatics analysis of missense mutations, and to assess the usefulness and reliability of the methods available (I-IV). This lead to the initiation of the development of a pipeline for the prediction of the effects of missense variants (IV).
3) Develop a new database for SH2 domain mutations (III)
4) The reliability of the methods for the prediction of missense variant pathogenicity was questioned in (II), which lead to the hypothesis that the reliability and prediction power of these methods may be insufficient to be used for the determination of possible pathogenicity of a variant. To test this hypothesis and facilitate the choice of methods to be used in the future, the performance of these methods was evaluated (V).

4. Materials and methods

4.1 Analyses of the effects of missense mutations (I-III)

4.1.1 Databases

Mutation data and amino acid sequences for the analysed proteins were obtained from IDbases (Piirilä et al. 2006) and SH2base (Lappalainen et al. 2008). The SH2base (http://bioinf.uta.fi/SH2base) was constructed by the MUTbase program suite (Riikonen and Vihinen 1999). All the mutations in human SH2 domains were found in literature and database searches. Individual locus specific databases were built for genes not included in the IDbases (*RASA1*, http://bioinf.uta.fi/RASA1base; and *PIK3R1*, http://bioinf.uta.fi/PIK3R1base). Information about the proteins, such as domain boundaries, functional sites and post-translational modification sites was extracted from literature and the UniProt database. Protein family information was obtained from the Pfam database (Finn et al. 2010). Three-dimensional structural data was obtained from the Protein Data Bank (PDB) (Berman et al. 2000).

4.1.2 Multiple sequence alignments and sequence conservation

Homologous sequences were retrieved by PSI-BLAST (Altschul et al. 1997), or extracted from the Pfam database (Finn et al.). Multiple sequence alignments were made with CLUSTALW (Thompson et al. 1994) and 3DCoffee (O'Sullivan et al. 2004). Ready-made sequence alignments for protein domains were extracted from Pfam. Alignments were visualised and the degree of evolutionary conservation for sequence positions was studied using MultiDisp (Riikonen and Vihinen, submitted), ProCon (Shen and Vihinen 2004) and ConSeq (Berezin et al. 2004). Conservation indices were calculated by AL2CO (Pei and Grishin 2001) and ConSurf (Glaser et al. 2003).

4.1.3 Disorder prediction

Structural disorder in proteins and the effect of missense mutation on disorder propensities were studied by the programs PONDR (Romero et al. 1997), DisEMBL (Linding et al. 2003a), GlobPlot (Linding et al. 2003b), DISOPRED (Ward et al. 2004), IUPred (Dosztányi et al. 2005), DRIP-PRED (http://www.sbc.su.se/~maccallr/disorder/) and Ronn (Yang et al. 2005). The effect of mutations was studied by comparing profiles for the wild-type sequences to those for each mutated sequence.

4.1.4 Aggregation prediction

The effects of mutations on protein aggregation propensities were predicted by the program TANGO (Fernandez-Escamilla et al. 2004) and calculations presented by Chiti (Chiti et al. 2003), for which the α-helical propensities were calculated by the program AGADIR (Muñoz and Serrano 1997).

4.1.5 Pathogenicity predictors

The predicted pathogenicities of mutations were studied by SNPs3D (Yue et al. 2006), SIFT (Ng and Henikoff 2001), PolyPhen (Sunyaev et al. 2001) and Pmut (Ferrer-Costa et al. 2005).

4.1.6 Stability prediction

The programs SCPred (Dosztányi et al. 1997; Dosztányi et al. 2003b), SCide (Dosztányi et al. 2003a; Dosztányi et al. 2003b), Sride (Gromiha et al. 2004), PoPMuSic (Sunyaev et al. 2001), FoldX (Schymkowitz et al. 2005) and DMUTANT (Zhou and Zhou 2002) were used for studying the effects of missense mutations on protein stability.

4.1.7 Structural analyses

Structural analyses were based on the crystal or solution structures obtained from PDB (Berman et al. 2000). Secondary structure boundaries were determined by the program STRIDE (Frishman and Argos 1995). The structures were visualised and mutated proteins modelled by PyMOL, version 0.99 (DeLano 2002). Hydrogen atoms were added to modelled structures by Reduce (Word et al. 1999), or Insight II (Accelrys), and mutant amino acid side chain χ angles were rotated at intervals of 10° by the Autobondrot function in PROBE 2.80 (Word et al. 2000). The rotatable side chains were created with PREKIN 5.93 (Word et al. 2000). The best fitting rotamers (judged by PROBE score) were selected and modelled on corresponding wild type structures. The threshold PROBE score for an acceptable conformation was -1.0, allowing for small perturbations in the structure (Lovell et al. 2000). Mutant structures were validated by the MolProbity server (Lovell et al. 2003). Models were then analysed for van der Waals effects and electrostatics using the programs PROBE and MAGE (Word et al. 2000), or PyMOL. The changes in contacts between residues in mutated structures versus wild type structures were studied by the programs WHAT IF (Vriend 1990), RankViaContact (Shen and Vihinen 2003), CSU (Sobolev et al. 1999) and MolProbity (Lovell et al. 2003), and visualised by KiNG (Lovell et al. 2003). Contact surfaces, as well as solvent accessible surfaces, were calculated with CSU (Sobolev et al. 1999). Electrostatic surface potentials were calculated and visualised by PyMOL.

4.2 Evaluation of the performance of prediction methods (V)

4.2.1 Datasets

The mutation dataset (34778 cases) was extracted from the PhenCode database (Giardine et al. 2007) and the dataset of neutral variants (32619 cases) from dbSNP (Sherry et al. 2001).

4.2.2 Prediction methods

The effects of disease-causing mutations (positive dataset) and SNPs (negative dataset) were predicted by the programs MutPred (Li et al., 2009) nsSNPAnalyzer (Bao et al. 2005), Panther (Thomas et al. 2003), PhD-SNP (Capriotti et al. 2006), Pmut (Ferrer-Costa et al. 2005), PolyPhen (Ramensky et al. 2002), SIFT (Ng and Henikoff 2001), SNAP (Bromberg and Rost 2007), and SNPs&GO (Calabrese et al. 2009). Default parameters were applied for all programs. Only the provided binary prediction (pathogenic/neutral) was taken into consideration from the output of programs, or numerical results were converted into binary predictions according to guidelines given by the authors of methods.

4.2.3 Structure coordinates and analysis

The 3D structure coordinates of proteins were obtained from PDB. Secondary structural information and solvent accessible surface area (SAS) values for each mutation site were assigned by the program STRIDE (Frishman and Argos 1995). Residues with SAS of <10% were classified as buried and with SAS >25% as exposed. The CATH database (Orengo et al. 1997) was used to group affected proteins according to secondary and tertiary structure types.

4.2.4 Calculating expected mutation values

Substitution statistics for both datasets were analysed by comparing frequencies of substitutions with the expected values calculated using the distribution of all amino acids in the datasets. For mutated residues, expected values were calculated with regard to their codon diversity, taking into account all possible amino acid substitutions.

4.2.5 Determining the significance of results

The χ^2 test was used to determine the significance of results and χ^2 was calculated as:

$$\chi^2 = \Sigma \frac{(f_o - f_e)^2}{f_e}$$

where f_o is the observed frequency and f_e is the expected frequency for an amino acid. P-values were estimated in a one-tailed fashion.

4.2.6 Statistical methods for the evaluation of prediction performance

The quality of predictions was described by six parameters: accuracy, precision, specificity, sensitivity, negative predictive value (NPV) and Matthews correlation coefficient (MCC). In the following equations, *tp, tn, fp,* and *fn* refer to the number of true positives, true negatives, false positives and false negatives, respectively.

$$\text{Accuracy} = \frac{tp + tn}{tp + tn + fp + fn}$$

$$\text{Precision} = \frac{tp}{tp + fp}$$

$$\text{Specificity} = \frac{tn}{fp + tn}$$

$$\text{Sensitivity} = \frac{tp}{tp + fn}$$

$$\text{NPV} = \frac{tn}{tn + fn}$$

$$\text{MCC} = \frac{tp \times tn - fn \times fp}{\sqrt{(tp + fn)(tp + fp)(tn + fn)(tn + fp)}}$$

Pearson correlations between program outputs were calculated by counting the cases common to all programs and those predicted similarly.

5. Summary of the results

5.1 Procedure for bioinformatics analysis of mutations

We developed a procedure for the bioinformatics analysis of mutations based on the experience in our group and previous studies on the effects of missense mutations (IV). The procedure was implemented and refined in the studies (I, II, III) and discussed in (IV). A schematic representation of the procedure is presented in Figure 3.

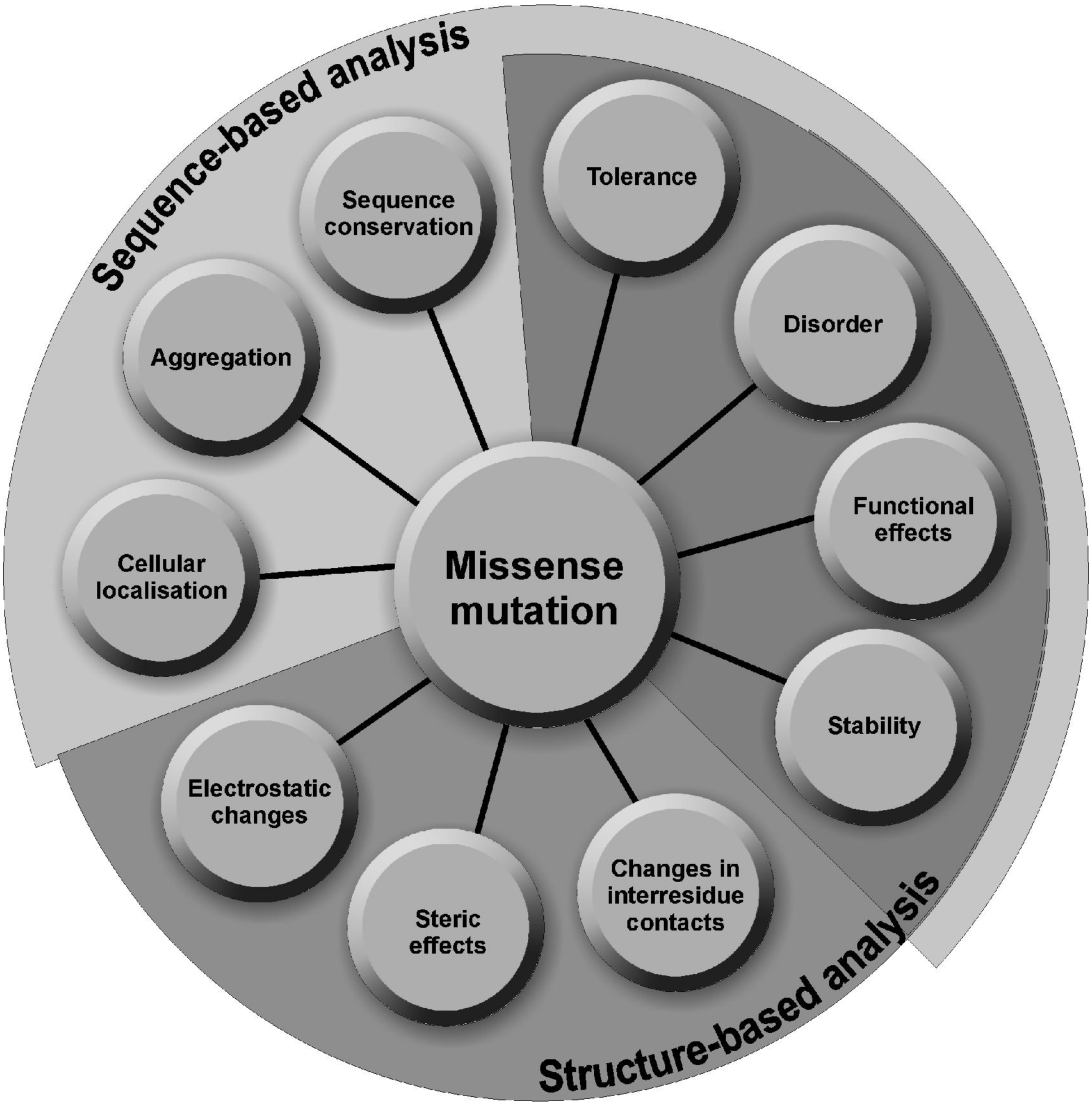


Figure 3. Groups of methods employed in the mutation analysis procedure. The approach is divided into sequence- and structure-based parts, which partly overlap.

5.2 Databases

The individual mutation databases for each of the diseases were updated for studies I-III. In addition, a new database for disease-causing mutations in SH2 domains, called SH2base, was created (III). In addition to mutation data, the SH2base contains information about SH2 domains and diseases caused by mutations within them. New locus specific databases were built for genes *RASA1* and *PIK3R1*, along with the development of the SH2base.

5.3 The molecular mechanisms of a selection of genetic diseases

In this study, we implemented the mutation analysis procedure in elucidating the molecular level effects of missense mutations causing cyclic and congenital neutropenia (I), XHIGM (II), XLA (III), XLP (III), SCID (III), NS (III), JMML (III), severe insulin deficiency (III), BCC (III), STAT1 deficiency (III), and growth hormone insensitivity with immunodeficiency (III). Features of the diseases are summarised in Table 1. For the diseases in the original communication (III), only those mutations affecting Src homology 2 (SH2) domains were analysed. Each known mutation was analysed individually for all possible effects it might have on protein structure or function. Missense mutations have diverse and parallel effects on protein structure and function and in many cases, features associated with a mutation overlap. Consequently, many missense mutations are predicted to have more than one effect in the analysis, which is why percentages in the following do not necessarily equal 100. The detailed results considering each disease and mutation are found in the original communications (I-III).

Table 1. Diseases analysed.

Gene	Protein	Disease	OMIM	Inheritance	Mutations[a]	Phenotypes
ELA2	HNE	Neutropenia, cyclic and congenital	162800, 130130	Autosomal dominant	32	Decreased or oscillating levels of neutrophils in the blood, recurrent infections
CD40L	CD40L	X-linked hyper-IgM syndrome (XHIGM)	308230	X-linked	35	Defective Ig class switching, recurrent infections.
BTK	BTK	X-linked agammaglobulinemia (XLA)	300300	X-linked	32	Hypogammaglobulinemia, antibody deficiency, recurrent infections
SH2D1A	SH2D1A	X-linked lymphoproliferative disease (XLP)	308240	X-linked	25	Mononucleosis, B cell lymphomas, dysgammaglobulinemia
ZAP70	ZAP-70	Severe combined immunodeficiency (SCID)	600802	Autosomal recessive	1	Severe pulmonary infection, chronic infections
PTPN11	SHP-2	Noonan syndrome (NS)	163955	Autosomal dominant	21	Short stature, facial dysmorphia, congenital heart defects
PTPN11	SHP-2	Juvenile myelomonocytic leukaemia (JMML)	607785	Not available	18	Myelodysplastic syndrome, leukaemia
PIK3R1	PI3-kinase	Severe insulin resistance	-	Not available	1	Hyperinsulinemia, diabetes mellitus at later stage
RASA1	RasGAP	Basal cell carcinoma (BCC)	605462	Sporadic	3	Clusters of basal cell carcinoma, tumours on the chest
STAT1	STAT1	STAT1 deficiency, complete	600555	Not available	1	Susceptibility to viral and intracellular bacterial infections
STAT5B	STAT5B	Growth hormone insensitivity with immunodeficiency	245590	Not available	1	Growth failure, recurrent infections

[a] Number of disease-causing missense mutations analysed in each study.

The missense mutations causing cyclic and congenital neutropenia were found to have a pronounced structural effect (I). Only three of the 32 amino acid substitutions could be accommodated in the structure and almost all substitutions were predicted to have effects on protein stability. Loss of protein stability leads to loss of enzymatic activity of HNE and possibly to the accumulation of the protein in cells. In contrast, the analysis of missense mutations in *CD40L* revealed the majority of mutations causing XHIGM have a mechanism of pathogenicity related to protein-protein interactions governing the function of the encoded protein (II). In SH2 domains, the pathogenic nature of mutations was determined to be related to the dysfunction in binding of proteins to their phosphopeptide ligands. This causes mutations to lead into defects in the corresponding signalling pathways in which the SH2 domain containing proteins serve a regulatory function (III). An exception among the SH2 domaining proteins was found, namely the protein SHP-2, which has a self-regulated enzymatic function distinct from the other SH2 domain containing proteins analysed. In SHP-2, disease-causing amino acid substitutions were not clustered in or around the ligand binding pocket, as in the other analysed proteins, but on the interdomain interface regulating enzyme function, causing dysregulation of the enzyme, thereby acting as gain-of-function mutations.

5.4 Sequence conservation and mutations affecting conserved positions

It is widely accepted that sequence positions conserved in evolution are commonly affected by disease-causing mutations, because these positions are generally important for the structure and function of proteins. There are three types of conservation that can be detected at the sequence level. The first type of conservation is invariance, where all amino acids occurring at the corresponding position in a multiple sequence alignment are the same. Conservation of physicochemical properties of amino acids, such as hydrophobicity or charge, is the second type. The third mechanism of conservation is covariation, where upon mutation, a compensating mutation occurs at another sequence position. When two or more positions in a protein family co-evolve, they are often involved in structural or functional networks in proteins.

We studied the level of conservation of positions that are affected by mutations in the aforementioned diseases, in order to find explanations for the pathogenicity of the mutations. Our results were in line with the general assumption that conserved positions are often affected by mutations. However, in many of the diseases we have studied here, a slight majority of missense mutations affect positions not conserved in evolution (II, III). Furthermore, there is variance in the nature of affected conserved positions in different proteins and diseases. In the TNF homology protein family, there are many evolutionarily invariant positions (Type I conservation) the majority of which (60 %) are affected by mutations in XHIGM (II). In contrast, there is only one invariant position in the trypsin homology family, in which there are no known mutations in human neutrophil elastase (HNE) that cause cyclic or congenital neutropenia (I). Only two mutations in CD40L affect positions with conserved physicochemical properties (Type II conservation) (II), whereas in HNE, many mutations affect positions where hydrophobicity is a conserved property (I). In CD40L, there are no mutations in covariant positions (Type III conservation) (II), but five co-varying positions are affected by mutations in HNE. In CD40L, although many conserved positions are mutated in XHIGM, the majority (63%) of all XHIGM-causing missense mutations affect positions that are not conserved in the TNF homology family of proteins (II). In HNE, the majority (60%) of the disease-causing mutations affect conserved positions (I). The differences in the sequence conservation patterns and the mutations affecting conserved positions in CD40L and HNE could be explained by the fact that in CD40L, a large fraction (37%) of disease-causing mutations affect residues involved in receptor binding and trimerisation of the protein (II); in other words, residues positioned at the surface of the protein. On the other hand, HNE mutations mostly affect buried, structure-maintaining residues (I). The variation in the results of sequence conservation analysis in these two proteins illustrates the importance of studying evolutionary conservation at different levels (Type I, Type II and Type III), instead of simply looking at the frequency of a particular amino acid at certain positions in the alignment. The different patterns of conservation can provide insight into possible disease mechanisms in different proteins. The picture is not always as simple, however, as illustrated by sequence conservation in SH2 domains and differences in the effects of mutations in BTK and SH2D1A. There is only one invariant residue in the SH2 domain family, but Type II and Type III conservation is evident. The fraction of conserved positions affected by mutations is approximately the same in both proteins (48 % and 46 %, respectively), but in BTK most of these positions are involved in protein function (phosphopeptide binding), while in SH2D1A the positions serve a structural role (III). In the SH2 domain family, the physicochemical properties of residues in the ligand-binding pocket affect protein function, whereas the properties of residues in the protein core have an effect on the structural integrity of the protein. Similarly, in the co-varying network of residues both structure-maintaining and function-related positions are found (III).

5.5 The effects of missense mutations on structural disorder

In studies of CD40L and HNE mutations, four and six methods for the prediction of disorder were used, respectively (I, II). The number was greater in (I) because the studies for the publication were done later than those for (II) and new methods were found. None of the mutations were predicted to cause disorder by all the methods used in either study. If the majority of methods predicted the mutation to increase disorder propensity in the protein, the mutation was considered likely to cause disorder. 25 % of the HNE mutations (I) and 23 % of the CD40L (II) mutations were likely to increase disorder in the protein. Mutations introducing a proline residue were over-represented in the cases predicted to cause disorder in proteins, which is in line with the well-known fact that prolines act as structural disruptors when placed the middle of secondary structure elements (Chou and Fasman 1974). All of the mutations predicted to cause disorder were found to have other structural effects on the corresponding proteins as well (I,II).

5.6 The effects of mutations on β aggregation

The results of β aggregation propensities of mutated proteins were controversial. In CD40L, five of the XHIGM-causing mutations were predicted to increase protein β aggregation propensity using the Chiti method, but only one of them was predicted to do so by the program TANGO (II). The results were parallel in HNE mutations: 4 missense mutations were predicted to cause protein β aggregation by the Chiti method, of which two agreed with the TANGO results. In addition, one mutation was predicted to increase aggregation propensity by TANGO, but not with the Chiti method (I). All mutations predicted to increase aggregation propensities of the corresponding proteins, were found to have additional effects on protein structure, predicted by other methods (I, II). It has been suggested that formation of native protein complexes and self-aggregation compete in the cell, as missense mutations affecting the interface or the stability of a protein complex, often lead to formation of toxic aggregates (Castillo and Ventura 2009). The missense mutations in homotrimer interfaces of CD40L were mostly not predicted to cause aggregation by the methods we used (II).

5.7 Mutations affecting protein structural stability

In general, the majority of missense mutations has been found to affect structurally important residues, causing loss of structural integrity or stability of the protein, rather than those residues directly involved in protein function (Mooney and Klein 2002; Wang and Moult 2001; Yue et al. 2005). Our results were mostly in line with this observation (I,II), with the exception of mutations in SH2 domains, where a slight majority (56%) of missense mutations affected residues involved in ligand binding and thus protein function (III). The mutations in SHP-2, the majority of which were found to be involved in regulation of protein activity and therefore assigned a functional role, are not included in the percentage. SHP-2 mutations illustrate the problem in artificial of categorising of mutations into groups, according to their effect. Disease-causing mutations are positioned at the interdomain interface in SHP-2, causing destabilisation of the inactive structure of the protein, thereby leading to defective regulation of its function (III). Depending on the point of view, these mutations could just as well be classified as structure destroying mutations.

Effects of mutations on the structural stability of proteins were studied by programs predicting stabilising residues in proteins and by programs evaluating mutation-induced stability changes in proteins (I, II). A substitution at a predicted stabilising residue was considered to destabilise structure. 56% of missense mutations in HNE were predicted to be structure destabilising by these methods. In addition to these methods, rotamer analysis and modelling of mutations on proteins were perfomed, in order to study whether the introduced side chain would fit into the structure, and if so, how the structure-maintaining contacts would change upon mutation (I, II, III). In HNE, only three of the 23 mutations screened could be adopted without significant rearrangements in the structure (I). Mutant residues that cannot fit into the structure lead to decreased stability of the protein, if they do not cause changes in folding or scaffolding of the protein, thereby causing the protein to be dysfunctional. Interestingly, an experimental study of a number of *ELA2* missense mutations has been made after we published (I) and the results show that the amino acid substitutions cause misfolding of the protein, leading to activation of unfolded protein response (UPR), followed by apoptosis of granulocyte precursor cells (Grenda et al., 2007). This result supports our prediction of extensive structure-destabilising effects of neutropenia-causing mutations. The majority of mutations in SH2D1A were found to have similar effects, as well as mutations in the STAT proteins (III). Of the CD40L mutations, about half the mutations were found to be unable to fit into the structure (II).

The contacts formed between amino acid side chains constitute the major structure maintaining and stabilising effects in folded proteins. As a consequence, alterations in the residue-residue contacts caused by mutant side chains were examined. Only those amino acid substitutions fitting into the structure according to the rotamer analysis were examined. Mutations causing loss of hydrophobic interactions, hydrogen bonds or disulphide bridges, or introducing charge in the protein core, thereby affecting the structural properties of proteins, were found to be very common (I-III).

The amino acid substitutions involving proline and glycine residues can be seen as a distinct group of conformation-affecting mutations because of the special characteristics of these residues. The fraction of mutations involving prolines was rather large among the neutropenia-causing mutations (34.4%) (I), compared to their occurrence in other analysed diseases (8.5% in XHIGM and 8.8% in the diseases caused by SH2 domain mutations) (II,III). This is in agreement with the general observation that missense mutations in *ELA2* lead to structural destabilisation of the protein (I), whereas in CD40L (II) and in the SH2 domain containing proteins (III), amino acid substitutions in general have a functional, rather than structural role.

5.8 Mutations affecting protein electrostatic surface potential and protein-protein interactions

Our results show the electrostatic surface potential of proteins is commonly affected by missense mutations. Many (37%) XHIGM-causing mutations were found to have an effect on the electrostatic surface potential of the protein, mostly causing the surface charge become more negative than the wild type (II). Results were similar for the SH2 domains (III). These mutations destabilise or hinder the interactions of these proteins with other molecules, because the interactions are driven by the positive charge of the binding surface of the wild type protein, which attracts the negatively charged partner molecule (II, III). The specific contact-forming amino acid residues at the ligand-receptor interface, as well as the interacting surfaces in the biologically active CD40L trimer, are well documented in literature. Of the 35 XHIGM-causing mutations, 16 were found to affect these residues, thereby influencing the molecular interactions essential for proper protein function and CD40L quaternary structure. Altogether, 54% of the CD40L mutations had an effect on the molecular interfaces. Mutations destabilising interfaces were common in SHP-2 as well, as many as 19 of the 24 mutations were positioned at the surface between the N-SH2 and protein tyrosine phosphatase (PTP) domains (III). In SH2 domains in general, 34 of the 103 analysed mutations presumably disturb ligand binding. The proportion of functional mutations was substantially smaller in HNE than in other proteins; only six mutations were located at the substrate binding site and affect the specificity of the enzyme. However, one quarter of the neutropenia-causing mutations were predicted to cause changes in the electrostatic surface potentential of HNE, although the overall majority of the mutations had an explicit structural role (I).

5.9 Genotype-phenotype correlations

Genotype-phenotype (GP) correlations are particularly interesting as mutations in one gene can lead to different disease phenotypes. Cyclic and congenital neutropenia are caused by mutations in the same gene (I), and the severity of XLA varies among patients (III). Mutations in *PTPN11* lead to Noonan syndrome, JMML and sometimes both (III).

We were not able to elucidate clear GP correlations for the diseases studied, one reason being that the same molecular event can cause different phenotypes in different families, or the phenotype may even vary within certain kindreds. This shows GP correlations are complicated and depend on factors beyond our methodology. GP correlations for *PTPN11* have been suggested to be dependent on the severity of the mutation considering its effect on the level of gain-of-function (Kratz et al. 2005). Because JMML-causing mutations are sporadic, they could be phenotypically more severe than the ones found in NS patients. We found a number of NS-causing mutations could be considered to have mild effect on the protein product, but also many NS-associated mutations have a more pronounced effect. Furthermore, some amino acid substitutions predicted to have a mild effect, lead to JMML (III). Elucidation of GP correlations would require systematic statistical studies, and the existing datasets with approximately 30 mutations for each disease, are not sufficient.

5.10 Reliability of missense mutation pathogenicity predictors (V)

A subset of the evaluated methods developed was found to perform reasonably well in terms of accuracy, precision (or positive predictive value, PPV), specificity, sensitivity, negative predictive value (NPV), and Matthews correlation (MCC).

According to our results, the overall best performing methods were SNPs&GO (Calabrese et al. 2009), SNAP (Bromberg and Rost 2007) and PolyPhen (Ramensky et al. 2002). SNPs&GO achieved an accuracy of 0.67, precision of 0.83, specificity of 0.91, and MCC of 0.39. In terms of sensitivity and NPV, SNAP was the best method. The results indicate that more reliable prediction methods are needed.

The methods were found to be very diverse in terms of performance, and interestingly, no convergence was observed between methods based on similar principles. For example, SNAP and Pmut (Ferrer-Costa et al. 2005) are both neural network based predictors sharing similar structural and sequence homology based attributes in describing missense variants. However, SNAP was one of the best performing methods (MCC 0.33) while Pmut was among the worst, with MCC of 0.09.

There was no appreciable difference in accuracy or specificity for substitutions at different types of secondary structural elements, buried or exposed positions, or different types of substituted or substituting amino acid. Sensitivity and the MCC of predictions varied according to these descriptors. MCC values also varied significantly with CATH structural class. In general, results of the prediction programs do not correlate well and only a small subset of the cases was correctly predicted by all methods.

6. Discussion

We have used numerous methods for the prediction of several different mechanisms by which missense mutations may affect protein structure and function. There are, in principle, two types of questions that could be answered by using the methodology presented in this study. Firstly, one may want to know whether a certain coding variant is pathogenic or benign, that is, whether it causes some kind of negative effects on the structure or function of the protein product or not. Secondly, one may be interested in why a certain pathogenic variant causes disease (and why a neutral variant does not) and what kind of negative effects it causes at the protein level.

6.1 Pathogenic or not?

The methods to be used should be chosen carefully depending on the question one wants to find an answer for. For example, the results from disorder prediction methods will not address the general question as to whether a missense variant is pathogenic or benign, but may be useful in analysing the molecular level mechanisms of pathogenicity for a mutation. Likewise, one would intuitively assume the results from the pathogenicity prediction methods would not be of any use when analysing the effects of a known disease-causing mutation. These methods, however, are based on different aspects and parameters describing pathogenicity, and could thus provide clues on the molecular level effects of a mutation. Furthermore, many pathogenicity prediction methods provide valuable information about the basis for the result; particularly those including structural information and/or database annotations in the prediction (see IV and V for the detailed description of individual methods). When addressing the question of whether a variant with unknown phenotypic effect is pathogenic or not, why follow our approach and use a multitude of different methods, when using just one of the pathogenicity predictors can provide a simple solution? Firstly, none of the methods is perfectly reliable, and secondly, none of them takes into account as many possible mechanisms of pathogenicity as our protocol (IV, V). Using a multiple method approach, such as our protocol, provides insight into the putative disease mechanism of the variant, beyond the pathogenic-or-not prediction methods. We find that by using multiple methods, based on different parameters and looking at various properties of mutations, probably more sophisticated and reliable results can be achieved. The finding that those missense variant pathogenicity prediction methods that include several descriptors of the variant in the prediction, perform better than those describing missense mutations in a more simple way also supports this view (V).

6.2 Sequence analysis

We used experimentally defined structures as a basis for interpretation of the effects of mutations (I-IV), but modelled structures can be used as well, when a structure has not yet been solved (Khan and Vihinen 2009). A subset of the methods in our approach is based on protein amino acid sequence only (Figure 3), and these can be employed in the absence of structural information. Sequence-based methods for identifying deleterious mutations have been shown to be as reliable as those methods incorporating structural data, in some studies (Ng and Henikoff 2001; Saunders and Baker 2002), but other studies have contradicted this view (Bromberg and Rost 2007; Krishnan and Westhead 2003). However, when studying the molecular level mechanisms of disease, structure-based methods can provide valuable information above and beyond sequence level.

For sequence-based approaches, there are (usually) plenty of sequences available, but the choice of sequences used in multiple sequence alignment (MSA) can drastically affect prediction accuracy. While an optimal alignment should use only orthologues with the same function, paralogues can reduce prediction accuracy. This is because a change in function would yield different conservation pressures in different regions of the protein. Many methods presented in I-V make automatic BLAST searches to generate input MSAs, which may impose problems on the quality of the MSA. In addition to possibly incorporating paralogue sequences in the

MSA, a common problem in running BLAST searches, is that there is typically a variable amount of redundancy in the set of sequence hits because multipe versions of the same sequence or mutant sequences may appear in the result. Duplicate sequences incorporated in the MSA may distort the sequence conservation profile. There are numerous methods available for multiple sequence alignment and deciding which would produce the most accurate results is a difficult task, as individual methods each have their specific strengths and weaknesses (Ahola et al. 2006, 2008) (discussed in IV). One approach would be to use a service that runs several MSA methods and combines them into a single model, such as the M-Coffee server (Moretti et al. 2007). We circumvented the problem in our analyses (I-III) by obtaining ready-made curated MSAs from Pfam (Bateman et al. 2004; Finn et al. 2010; Sonnhammer et al. 1997), when available.

6.3 2D structure and disorder

Secondary structure predictors can be useful in predicting the structural framework in which missense mutations are found, in order to make hypotheses about their role. These methods have not been used in (I-III), however, because the structures of the proteins analysed were solved experimentally. According to our experience, the existing secondary structure prediction methods are not sensitive enough to detect missense mutation-induced changes in 2D structural elements.

This observation leads to the question of whether disorder prediction methods are sensitive enough, as disorder essentially refers to the lack of an ordered secondary structure. Furthermore, these methods have not been developed for the analysis of missense mutations on disorder propensity, but to predict whether a protein of unknown structure is intrinsically disordered or contains disordered elements. The performances of methods for disorder prediction have been evaluated and reported in context of their development, but performance in the analysis of variants has not been studied. In our studies (I-II) the results of different disorder prediction methods were inconsistent, which most probably results from different methods implementing diverse parameters for prediction, and even the concept of disorder lacks a categorical definition (discussed further in IV). We strongly feel the accuracy of these methods should be evaluated, firstly because we do not know whether they are suitable for analysis of missense mutations considering their sensitivity and secondly, the abundance of existing methods based on different parameters and training sets. This type of an analysis falls, however, outside the scope of this thesis. Until the disorder prediction methods have been evaluated, our view is that they can be used qualitatively for generating hypotheses about possible mechanisms of mutation pathogenicity, but the results should be interpreted critically in light of the outcome of other methods. For example, if a mutation is predicted to increase disorder by several disorder methods (judged by a majority vote) and it is found to have conformational or stability decreasing effects as well, it could be hypothesised that the mutation has secondary/tertiary structure-destabilising and disorder-increasing effects, causing dysfunctionality of the protein and thereby disease. Alternatively, if all other methods failed to provide any explanation for the molecular mechanism of a mutation, the outcome of the disorder methods could be used in speculating the possible structural effects of that mutation.

6.4 Structure analysis

The structural effects of missense mutations have been studied by modelling amino acid substitution onto experimentally solved structures (I-III). With regards to protein folding, can it be assumed that despite the amino acid substitution, the protein would fold into a three-dimensional structure identical to the wild-type protein and structural details (such as interresidue contacts the substituting residue forms) could be studied as we propose in our procedure? Before modelling, we analysed whether it is theoretically possible to accommodate the substituting residue into the existing structure. If not, those cases were left out from further steps of structure analysis, because our method cannot predict the necessary structural changes in the vicinity of the substitution. In other cases, the theoretically best-fitting conformation of the residue side chain was employed. We acknowledge that there are factors this theoretical approach does not take into account, but useful and correct predictions about the effects of mutations have been reached by molecular modelling, even without an experimental structure for the wild type protein (Khan and Vihinen 2009). The prediction of mutational effects is even more difficult when using modelled proteins because even with high sequence homology between the template and target proteins the positions of residue side chains cannot be precisely computed. More detailed information about the effects of amino acid substitutions on protein structure, folding and dynamics, could be attained by molecular dynamics simulations, which are computationally heavy. In this instance, cases to be analysed should be selected carefully by applying a procedure for prioritisation of the most interesting cases – a procedure such as the one presented in this study.

6.5 Structural and functional mutations

The classification of mutational effects into structure-perturbing or function-abolishing categories employed in (I-III) is controversial, considering the intimate relationship between protein structure and function. For example, should amino acid substitutions in interdomain or intermolecular interfaces be classified as structural or functional mutations? Substitutions affecting the interdomain interface in the SHP-2 protein are predicted to lead to constant activation of the enzyme by abolishing the regulatory function of the SH2 domain when interacting with the phosphatase domain (III). Thus, these mutations have been classified as directly related to function, although they could be viewed as structure-affecting (destabilisation of quaternary structure) as well. Similarly, in RasGAP the R398L substitution is predicted to affect the structure of the phosphotyrosine-binding pocket, through loss of hydrogen bonding with other residues involved in pocket formation (III). Structural alterations in the binding pocket are predicted to lead to changes in ligand binding and thereby function, so the mutation is classified as directly related to protein function. In HNE, a neutropenia-causing mutation causes the substitution of cysteine 71 by arginine. The residue is located at a site determining specificity of the protein, thus the substitution is predicted to alter the specificity of the protein and the mutation is classified to be functional (I). However, C71 also forms a disulphide bond with C55, the breakage of which probably leads to destabilisation of the protein. Thus, amino acid substitutions can have multiple effects on protein structure and function, and the prediction of these effects sometimes produces overlapping results. In cases like this, our method cannot determine which mechanism is the primary causative effect of the disease phenotype. The categorisation of mutations into functional or structural classes according to effects at the protein level depends on the point of view and therefore this categorisation is a relative, rather than absolute statement of the nature of the mutation.

6.6 Reliability of the methods used

The performance and accuracy of individual methods has been evaluated and reported by the developers. In order to assess the reliability of the results obtained by our approach however, fair and comprehensive evaluation of the reliability of each prediction method in the procedure should be performed, followed by experimental verification of the predictive findings. Furthermore, there is a plethora of bioinformatics methods available for prediction of different aspects of missense variant effects (IV) and the number of methods employed in the approach should be diminished for efficiency. Evaluation of methods is needed in order to be able to select the most reliable ones. We started the evaluation task by studying reliability of missense variant pathogenicity prediction methods in (V) employing a similar approach as in (Khan and Vihinen 2010), where stability prediction methods were evaluated. Some mutational effects we have predicted in (I) have been studied experimentally (Grenda et al., 2007), providing support for our hypotheses and success of our procedure for the analysis of molecular effects of mutations (I-IV).

6.7 Analysis of mutational effects

Study of the effects of missense mutations at the protein level can give useful and interesting insight into the molecular basis of hereditary diseases as presented in (I-IV). The mutation analysis approach developed in this study is, to our knowledge, the most comprehensive in terms of the number of methods used. Our approach is also novel in terms of employing a multitude of viewpoints into protein-level consequences of a missense mutation. In order to attain more accurate results and even more comprehensive insights about mechanisms of pathogenicity of mutations, the analysis could be complemented with studies of effects of mutations at the DNA and RNA level. Our method does not take into account any mutational effects on the expression of the gene, but is based on the assumption that any mutation can have an effect at the protein level. Furthermore, given the redundancy of several cellular pathways, the molecular consequences of missense mutations should be viewed in the cellular context. A cellular-level view of the mechanisms of pathogenicity could be achieved by applying a systems biology approach.

We feel evaluation and verification of the bioinformatics approach for the analysis of mutational effects is of utmost importance, so that computational prediction could obtain a more established role in the study of human hereditary disease, accelerating for example, the identification of pharmaceutical targets for relevant treatments. It should be noted that despite being useful in providing information about the nature of mutations as such, bioinformatics analyses could also be helpful in guiding the design of further experimental research.

7. Summary and conclusions

In this study we have developed a procedure for analysis of the effects of nsSNPs and missense mutations on proteins. The analysis has provided interesting insights into molecular level mechanisms of pathogenicity of disease-causing mutations in a number of hereditary diseases, comprising immunodeficiencies and diseases caused by mutations in SH2 domains. Defects in HNE structural scaffolding and stability were the cause for pathogenicity of missense mutations in cyclic and congenital neutropenia. In X-linked Hyper IgM syndrome, mutations in CD40L have wide-ranging effects, primarily affecting the quaternary assembly of the protein and interactions it forms with its receptor. In SH2 domain containing proteins, pathogenicity of the mutations was mainly caused by effects on ligand binding and subsequent protein activation and therefore on downstream cellular signalling pathways. A special case among the SH2 domain proteins was found, SHP-2, in which the disease-causing amino acid substitutions were found to have a gain-of-function effect by disrupting the interdomain interface involved in suppressing activity of the protein. The genome-wide analysis of pathogenic SH2 domain mutations also lead to the development of a mutation database dedicated for disease-causing SH2 domain mutations, aiding further research on SH2 domains and pathologies associated with them.

In addition to elucidating the molecular basis of hereditary diseases, gaining knowledge about possible disease associations of nsSNPs detected in large-scale sequencing efforts is another major research problem. Vast amounts of variation data are being produced, providing invaluable opportunities to gain knowledge about genetic determinants of phenotypic variation and disease susceptibility in humans. For the time being, knowledge of the effects of genetic variants on phenotypes is lagging behind data accumulation. Many pathogenic variants have been identified (Hamosh et al. 2005; Minoshima et al. 2001), but the number of known disease-causing variations is small compared to the number of known polymorphisms and it is still unclear which polymorphisms have biological effects.

To understand the molecular basis of effects of human genetic variations on phenotype, predictive analysis of the effects of polymorphisms on gene function in all human genes is needed. There are several methods available for the prediction of missense variant pathogenicity and their performance was evaluated. The results indicate that although some of the methods reach reasonably accurate predictions more reliable methods should be developed in order to meet the need for efficiency and reliability in mutation research.

The mutation analysis procedure developed in this study is a comprehensive approach for the predictive analysis of the wide-ranging mutational effects at the protein level. The procedure has been streamlined during the course of this study and evaluation of the performance of its constituent programs has been started. A mutation effect metaserver is currently being developed based on the results and experiences gained herein. The novel approach for the predictive analysis of the effects of missense variants could be used in drawing hypotheses on the molecular causes of disease, with prioritisation of cases with unknown effects for further study, guiding the course of experimental studies, or aiding the design of proteins with novel properties for biotechnological applications.

8. Acknowledgements

This thesis work was carried out in the Bioinformatics group at the Institute of Medical Technology, University of Tampere. The research was funded by the Tampere Graduate School in Biomedicine and Biotechnology (TGSBB) and the Sigrid Jusélius Foundation.

I am more than thankful to my supervisor Prof. Mauno Vihinen for all these years I have worked with him. Mauno inspired me to start working in the field in the first place, and he has taught me a great deal about research and scientific thinking. I am grateful for his patience and flexibility and most of all for his constant encouragement and motivation. I am thanking Mauno for always believing in me and giving me the opportunity to grow to be a scientist. With his accomplished guidance it has been possible to complete this thesis.

I am also grateful to my thesis committee members, Jarkko Valjakka and Bairong Shen, for their advice and support.

The official reviewers of my thesis, Professor Rita Casadio and Docent Heikki Lehväslaiho, are gratefully appreciated for their careful evaluation of the manuscript and valuable comments and suggestions. I thank Professor Anthony J. Brookes for doing me the honour of being my opponent. I also wish to thank Eloise Kok for the language revision of this thesis.

During my years in the Bioinformatics group I have had the honour of working with many wonderful and talented people, who are warmly thanked for generating such a friendly, helpful and welcoming atmosphere in the group. Especially I would like to thank Jukka Lehtiniemi who gave the final touch for almost all figures in my publications, posters, presentations and in this thesis, and also helped me with issues in web publishing. I am thanking him for his patience and helpfulness, and also for his magnificent sense of humour. I am also thanking Jouni Väliaho for always being there to assist, and especially for helping me so much in the beginning of my project. I am grateful to Martti Tolvanen for teaching me a great deal about molecular visualization, to Hannu Korhonen for helping me out with crashing

computers, and to Ayodeji Olatubosun for scripting assistance and his admirable patience and devotion to science. Sofia Khan is acknowledged for the peer support during the final stages of our theses. I would also like to express my deepest gratitude to the other past and present group members for generating the group spirit, sharing everyday research life, and for the good times at and outside work.

I wish to thank my amazing friends for inspiration, support and sharing, and my family for all their help and encouragement.

Tampere, April 2010

Janita Thusberg

9. References

Abkevich V, Zharkikh A, Deffenbaugh AM, Frank D, Chen Y, Shattuck D, Skolnick MH, Gutin A, and Tavtigian SV (2004): Analysis of missense variation in human BRCA1 in the context of interspecific sequence variation. J Med Genet 41: 492-507.

Ahola V, Aittokallio T, Vihinen M, and Uusipaikka E (2006): A statistical score for assessing the quality of multiple sequence alignments. BMC Bioinformatics 7: 484.

Ahola V, Aittokallio T, Vihinen M, and Uusipaikka E (2008): Model-based prediction of sequence alignment quality. Bioinformatics 24: 2165-71.

Allanson JE (1987): Noonan syndrome. J Med Genet 24: 9-13.

Allen RC, Armitage RJ, Conley ME, Rosenblatt H, Jenkins NA, Copeland NG, Bedell MA, Edelhoff S, Disteche CM, Simoneaux DK, Fanslow WC, Belmont J, and Spriggs MK (1993): CD40 ligand gene defects responsible for X-linked hyper-IgM syndrome. Science 259: 990-3.

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, and Lipman DJ (1997): Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25: 3389-402.

Aly AM, Higuchi M, Kasper CK, Kazazian HH, Jr., Antonarakis SE, and Hoyer LW (1992): Hemophilia A due to mutations that create new N-glycosylation sites. Proc Natl Acad Sci U S A 89: 4933-7.

Arnold A, Horst SA, Gardella TJ, Baba H, Levine MA, and Kronenberg HM (1990): Mutation of the signal peptide-encoding region of the preproparathyroid hormone gene in familial isolated hypoparathyroidism. J Clin Invest 86: 1084-7.

Arpaia E, Shahar M, Dadi H, Cohen A, and Roifman CM (1994): Defective T cell receptor signaling and CD8+ thymic selection in humans lacking zap-70 kinase. Cell 76: 947-58.

Bach-Gansmo ET, Halvorsen S, Godal HC, and Skjønsberg OH (1996): D-dimers are degraded by human neutrophil elastase. Thromb Res 82: 177-86.

Baker D and Sali A (2001): Protein structure prediction and structural genomics. Science 294: 93-6.

Bao L and Cui Y (2005): Prediction of the phenotypic effects of non-synonymous single nucleotide polymorphisms using structural and evolutionary information. Bioinformatics 21: 2185-90.

Bao L, Zhou M, and Cui Y (2005): nsSNPAnalyzer: identifying disease-associated nonsynonymous single nucleotide polymorphisms. Nucleic Acids Res 33: W480-2.

Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, Khanna A, Marshall M, Moxon S, Sonnhammer EL, Studholme DJ, Yeats C, and Eddy SR (2004): The Pfam protein families database. Nucleic Acids Res 32: D138-41.

Baynes KC, Beeton CA, Panayotou G, Stein R, Soos M, Hansen T, Simpson H, O'Rahilly S, Shepherd PR, and Whitehead JP (2000): Natural variants of human p85α phosphoinositide 3-kinase in severe insulin resistance: a novel variant with impaired insulin-stimulated lipid kinase activity. Diabetologia 43: 321-31.

Bentires-Alj M, Paez JG, David FS, Keilhack H, Halmos B, Naoki K, Maris JM, Richardson A, Bardelli A, Sugarbaker DJ, Richards WG, Du J, Girard L, Minna JD, Loh ML, Fisher DE, Velculescu VE, Vogelstein B, Meyerson M, Sellers WR, and Neel BG (2004): Activating mutations of the noonan syndrome-associated SHP2/PTPN11 gene in human solid tumors and adult acute myelogenous leukemia. Cancer Res 64: 8816-20.

Benzeno S, Lu F, Guo M, Barbash O, Zhang F, Herman JG, Klein PS, Rustgi A, and Diehl JA (2006): Identification of mutations that disrupt phosphorylation-dependent nuclear export of cyclin D1. Oncogene 25: 6291-303.

Berezin C, Glaser F, Rosenberg J, Paz I, Pupko T, Fariselli P, Casadio R, and Ben-Tal N (2004): ConSeq: the identification of functionally and structurally important residues in protein sequences. Bioinformatics 20: 1322-4.

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, and Bourne PE (2000): The Protein Data Bank. Nucleic Acids Res 28: 235-42.

Betz SF (1993): Disulfide bonds and the stability of globular proteins. Protein Sci 2: 1551-8.

Bodak N, Queille S, Avril MF, Bouadjar B, Drougard C, Sarasin A, and Daya-Grosjean L (1999): High levels of patched gene mutations in basal-cell carcinomas from patients with xeroderma pigmentosum. Proc Natl Acad Sci U S A 96: 5117-22.

Bode AM and Dong Z (2004): Post-translational modification of p53 in tumorigenesis. Nat Rev Cancer 4: 793-805.

Bogan AA and Thorn KS (1998): Anatomy of hot spots in protein interfaces. J Mol Biol 280: 1-9.

Bourhis JM, Canard B, and Longhi S (2007): Predicting protein disorder and induced folding: from theoretical principles to practical applications. Curr Protein Pept Sci 8: 135-49.

Boxer LA and Morganroth ML (1987): Neutrophil function disorders. Dis Mon 33: 681-780.

Briscoe AD, Gaur C, and Kumar S (2004): The spectrum of human rhodopsin disease mutations through the lens of interspecific variation. Gene 332: 107-18.

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

Brown CJ, Srinivasan D, Jun LH, Coomber D, Verma CS, and Lane DP (2008): The electrostatic surface of MDM2 modulates the specificity of its interaction with phosphorylated and unphosphorylated p53 peptides. Cell Cycle 7: 608-10.

Bruton OC (1952): Agammaglobulinemia. Pediatrics 9: 722-8.

Brylinski M and Skolnick J (2008): A threading-based method (FINDSITE) for ligand-binding site prediction and functional annotation. Proc Natl Acad Sci U S A. 105: 129-34.

Bucciantini M, Calloni G, Chiti F, Formigli L, Nosi D, Dobson CM, and Stefani M (2004): Prefibrillar amyloid protein aggregates share common features of cytotoxicity. J Biol Chem 279: 31374-82.

Buckle AM, Cramer P, and Fersht AR (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: 4298-305.

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

Capra JA and Singh M (2007): Predicting functionally important residues from sequence conservation. Bioinformatics 23: 1875-82.

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: 2729-34.

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, and Lander ES (1999). Characterization of single-nucleotide polymorphisms in coding regions of human genes. Nat Genet 22: 231-8.

Carvalho M, Pino MA, Karchin R, Beddor J, Godinho-Netto M, Mesquita RD, Rodarte RS, Vaz DC, Monteiro VA, Manoukian S, Colombo M, Ripamonti CB, Rosenquist R, Suthers G, Borg A, Radice P, Grist SA, Monteiro AN, and Billack B (2009): Analysis of a set of missense, frameshift, and in-frame deletion variants of BRCA1. Mutat Res 660: 1-11.

Carvalho MA, Marsillac SM, Karchin R, Manoukian S, Grist S, Swaby RF, Urmenyi TP, Rondinelli E, Silva R, Gayol L, Baumbach L, Sutphen R, Pickard-Brzosowicz JL, Nathanson KL, Sali A, Goldgar D, Couch FJ, Radice P, and Monteiro AN (2007): Determination of cancer risk associated with germ line BRCA1 missense variants by functional analysis. Cancer Res 67: 1494-501.

Casari G, Sander C, and Valencia A (1995): A method to predict functional residues in proteins. Nat Struct Biol 2: 171-8.

Castillo V and Ventura S (2009): Amyloidogenic regions and interaction surfaces overlap in globular proteins related to conformational diseases. PLoS Comput Biol 5: e1000476.

Cavallo A and Martin AC (2005): Mapping SNPs to protein sequence and structure data. Bioinformatics 21: 1443-50.

Cerdà-Costa N, Esteras-Chopo A, Avilés FX, Serrano L, and Villegas V (2007): Early kinetics of amyloid fibril formation reveals conformational reorganisation of initial aggregates. J Mol Biol 366: 1351-63.

Chan AC, Iwashima M, Turck CW, and Weiss A (1992): ZAP-70: a 70 kd protein-tyrosine kinase that associates with the TCR ζ chain. Cell 71: 649-62.

Chan AC, Kadlecek TA, Elder ME, Filipovich AH, Kuo WL, Iwashima M, Parslow TG, and Weiss A (1994): ZAP-70 deficiency in an autosomal recessive form of severe combined immunodeficiency. Science 264: 1599-601.

Chang YF, Imam JS, and Wilkinson MF (2007): The nonsense-mediated decay RNA surveillance pathway. Annu Rev Biochem 76: 51-74.

Chasman D and 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.

Chiti F, Stefani M, Taddei N, Ramponi G, and Dobson CM (2003): Rationalization of the effects of mutations on peptide and protein aggregation rates. Nature 424: 805-8.

Chiti F, Webster P, Taddei N, Clark A, Stefani M, Ramponi G, and Dobson CM (1999): Designing conditions for in vitro formation of amyloid protofilaments and fibrils. Proc Natl Acad Sci U S A 96: 3590-4.

Chou JY and Mansfield BC (2008): Mutations in the glucose-6-phosphatase-α (G6PC) gene that cause type Ia glycogen storage disease. Hum Mutat 29: 921-30.

Chou PY and Fasman GD (1974): Prediction of protein conformation. Biochemistry 13: 222-45.

Chung JL, Wang W and Bourne PE (2006): Exploiting sequence and structure homologs to identify protein-protein binding sites. Proteins 62: 630-40.

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

Coffey AJ, Brooksbank RA, Brandau O, Oohashi T, Howell GR, Bye JM, Cahn AP, Durham J, Heath P, Wray P, Pavitt R, Wilkinson J, Leversha M, Huckle E, Shaw-Smith CJ, Dunham A, Rhodes S, Schuster V, Porta G, Yin L, Serafini P, Sylla B, Zollo M, Franco B, Bolino A, Seri M, Lanyi A, Davis JR, Webster D, Harris A, Lenoir G, de St Basile G, Jones A, Behloradsky BH, Achatz H, Murken J, Fassler R, Sumegi J, Romeo G, Vaudin M, Ross MT, Meindl A, and Bentley DR (1998): Host response to EBV infection in X-linked lymphoproliferative disease results from mutations in an SH2-domain encoding gene. Nat Genet 20: 129-35.

Cohen P (2006): The twentieth century struggle to decipher insulin signalling. Nat Rev Mol Cell Biol 7: 867-73.

Dale DC, Person RE, Bolyard AA, Aprikyan AG, Bos C, Bonilla MA, Boxer LA, Kannourakis G, Zeidler C, Welte K, Benson KF, and Horwitz M (2000): Mutations in the gene encoding neutrophil elastase in congenital and cyclic neutropenia. Blood 96: 2317-22.

DeLano WL (2002): The PyMOL molecular graphics system. DeLano Scientific, San Carlos, CA. http://www.pymol.org.

Dobson CM (2004): Principles of protein folding, misfolding and aggregation. Semin Cell Dev Biol 15: 3-16.

Dosztányi Z, Csizmók V, Tompa P, and Simon I (2005): IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. Bioinformatics 21: 3433-4.

Dosztányi Z, Fiser A, and Simon I (1997): Stabilization centers in proteins: identification, characterization and predictions. J Mol Biol 272: 597-612.

Dosztányi Z, Magyar C, Tusnády G, and Simon I (2003a): SCide: identification of stabilization centers in proteins. Bioinformatics 19: 899-900

Dosztányi Z, Magyar C, Tusnády GE, Cserzo M, Fiser A, and Simon I (2003b): Servers for sequence-structure relationship analysis and prediction. Nucleic Acids Res 31: 3359-63.

Dosztányi Z, Sandor M, Tompa P, and Simon I (2007): Prediction of protein disorder at the domain level. Curr Protein Pept Sci 8: 161-71.

Dundas J, Ouyang Z, Tseng J, Binkowski A, Turpaz Y, and Liang J (2006): CASTp: computed atlas of surface topography of proteins with structural and topographical mapping of functionally annotated residues. Nucleic Acids Res 34: W116-8.

Dunker AK, Silman I, Uversky VN, and Sussman JL (2008): Function and structure of inherently disordered proteins. Curr Opin Struct Biol 18: 756-64.

Dupuis S, Jouanguy E, Al-Hajjar S, Fieschi C, Al-Mohsen IZ, Al-Jumaah S, Yang K, Chapgier A, Eidenschenk C, Eid P, Al Ghonaium A, Tufenkeji H, Frayha H, Al-Gazlan S, Al-Rayes H, Schreiber RD, Gresser I, and Casanova JL (2003): Impaired response to interferon-α/β and lethal viral disease in human STAT1 deficiency. Nat Genet 33: 388-91.

Eriksson AE, Baase WA, Zhang XJ, Heinz DW, Blaber M, Baldwin EP, and Matthews BW (1992): Response of a protein structure to cavity-creating mutations and its relation to the hydrophobic effect. Science 255: 178-83.

Esteras-Chopo A, Serrano L, and Lopez de la Paz M (2005): The amyloid stretch hypothesis: recruiting proteins toward the dark side. Proc Natl Acad Sci U S A 102: 16672-7.

Eswar N and Ramakrishnan C (2000): Deterministic features of side-chain main-chain hydrogen bonds in globular protein structures. Protein Eng 13: 227-38.

Facchiano A and Marabotti A (2010): Analysis of galactosemia-linked mutations of GALT enzyme using a computational biology approach. Protein Eng Des Sel 23: 103-13.

Fandrich M, Fletcher MA, and Dobson CM (2001): Amyloid fibrils from muscle myoglobin. Nature 410: 165-6.

Fernandez-Escamilla AM, Rousseau F, Schymkowitz J, and Serrano L (2004): Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins. Nat Biotechnol 22: 1302-6.

Ferrer-Costa C, Gelpí JL, Zamakola L, Parraga I, de la Cruz X, and Orozco M (2005): PMUT: a web-based tool for the annotation of pathological mutations on proteins. Bioinformatics 21: 3176-8.

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: 771-86.

Fingerhut A, Reutrakul S, Knuedeler SD, Moeller LC, Greenlee C, Refetoff S, and Janssen OE (2004): Partial deficiency of thyroxine-binding globulin-Allentown is due to a mutation in the signal peptide. J Clin Endocrinol Metab 89: 2477-83.

Finn RD, Mistry J, Tate J, Coggill P, Heger A, Pollington JE, Gavin OL, Gunasekaran P, Ceric G, Forslund K, Holm L, Sonnhammer EL, Eddy SR, and Bateman A (2010): The Pfam protein families database. Nucleic Acids Res 38: D211-22.

Fleming MA, Potter JD, Ramirez CJ, Ostrander GK, and Ostrander EA (2003): Understanding missense mutations in the BRCA1 gene: an evolutionary approach. Proc Natl Acad Sci U S A 100: 1151-6.

Forbes SA, Tang G, Bindal N, Bamford S, Dawson E, Cole C, Kok CY, Jia M, Ewing R, Menzies A, Teague JW, Stratton MR, and Futreal PA (2010): COSMIC (the Catalogue of Somatic Mutations in Cancer): a resource to investigate acquired mutations in human cancer. Nucleic Acids Res 38: D652-7.

Fowler SB, Poon S, Muff R, Chiti F, Dobson CM, and Zurdo J (2005): Rational design of aggregation-resistant bioactive peptides: reengineering human calcitonin. Proc Natl Acad Sci U S A 102: 10105-10.

Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA, Belmont JW, Boudreau A, Hardenbol P, Leal SM, Pasternak S, Wheeler DA, Willis TD, Yu F, Yang H, Zeng C, Gao Y, Hu H, Hu W, Li C, Lin W, Liu S, Pan H, Tang X, Wang J, Wang W, Yu J, Zhang B, Zhang Q, Zhao H, Zhou J, Gabriel SB, Barry R, Blumenstiel B, Camargo A, Defelice M, Faggart M, Goyette M, Gupta S, Moore J, Nguyen H, Onofrio RC, Parkin M, Roy J, Stahl E, Winchester E, Ziaugra L, Altshuler D, Shen Y, Yao Z, Huang W, Chu X, He Y, Jin L, Liu Y, Sun W, Wang H, Wang Y, Xiong X, Xu L, Waye MM, Tsui SK, Xue H, Wong JT, Galver LM, Fan JB, Gunderson K, Murray SS, Oliphant AR, Chee MS, Montpetit A, Chagnon F, Ferretti V, Leboeuf M, Olivier JF, Phillips MS, Roumy S, Sallée C, Verner A, Hudson TJ, Kwok PY, Cai D, Koboldt DC, Miller RD, Pawlikowska L, Taillon-Miller P, Xiao M, Tsui LC, Mak W, Song YQ, Tam PK, Nakamura Y, Kawaguchi T, Kitamoto T, Morizono T, Nagashima A, Ohnishi Y, Sekine A, Tanaka T, Tsunoda T, Deloukas P, Bird CP, Delgado M, Dermitzakis ET, Gwillliam R, Hunt S, Morrison J, Powell D, Stranger BE, Whittaker P, Bentley DR, Daly MJ, de Bakker PI, Barrett J, Chretien YR, Maller J, McCarroll S, Patterson N, Pe'er I, Price A, Purcell S, Richter DJ, Sabeti P, Saxena R, Schaffner SF, Sham PC, Varilly P, Altschuler D, Stein LD, Krishnan L, Smith AV, Tello-Ruiz MK, Thorisson GA, Chakravarti A, Chen PE, Cutler DJ, Kashuk CS, Lin S, Abecasis GR, Guan W, Li Y, Munro HM, Qin ZS, Thomas DJ, McVean G, Auton A, Bottolo L, Cardin N, Eyheramendy S, Freeman C, Marchini J, Myers S, Spencer C, Stephens M, Donnelly P, Cardon LR, Clarke G, Evans DM, Morris AP, Weir BS, Tsunoda T, Mullikin JC, Sherry ST, Feolo M, Skol A, Zhang H, Zeng C, Zhao H, Matsuda I, Fukushima Y, Macer DR, Suda E, Rotimi CN, Abedamowo CA, Ajayi I, Aniagwu T, Marshall PA, Nkwodimmah C, Royal CD, Leppert MF, Dixon M, Peiffer A, Qiu R, Kent A, Kato K, Niikawa N, Adewole IF, Knoppers BM, Foster MW, Clayton EW, Watkin J, Gibbs RA, Belmont JW, Muzny D, Nazareth L, Sodegren E, Weinstock GM, Wheeler DA, Yakub I, Gabriel SB, Onofrio RC, Richter DJ, Ziaugra L, Birren BW, Daly MJ, Altschuler D, Wilson RK, Fulton LL, Rogers J, Burton J, Carter NP, Clee CM, Griffiths M, Jones MC, McLay K, Plumb RW, Ross MT, Sims SK, Willey DL, Chen Z, Han H, Kang L, Godbout M, Wallenburg JC, L'Archevêque P, Bellemare G, Saeki K, Wang H, An D, Fu H, Li Q, Wang Z, Wang R, Holden AL, Brooks LD, McEwen JE, Guyer MS, Wang VO, Peterson JL, Shi M, Spiegel J, Sung LM, Zacharia LF, Collins FS, Kennedy K, Jamieson R and Stewart J (2007): A second generation human haplotype map of over 3.1 million SNPs. Nature 449: 851-61.

Friedman E, Gejman PV, Martin GA, and McCormick F (1993): Nonsense mutations in the C-terminal SH2 region of the GTPase activating protein (GAP) gene in human tumours. Nat Genet 5: 242-7.

Frischmeyer PA, van Hoof A, O'Donnell K, Guerrerio AL, Parker R, and Dietz HC (2002). An mRNA surveillance mechanism that eliminates transcripts lacking termination codons. Science 295: 2258-61.

Frishman D and Argos P (1995). Knowledge-based protein secondary structure assignment. Proteins 23: 566-79.

Fuleihan R, Ramesh N, and Geha RS (1993): Role of CD40-CD40-ligand interaction in Ig-isotype switching. Curr Opin Immunol 5: 963-7.

George Priya Doss C, Sudandiradoss C, Rajasekaran R, Purohit R, Ramanathan K, and Sethumadhavan R (2008): Identification and structural comparison of deleterious mutations in nsSNPs of ABL1 gene in chronic myeloid leukemia: a bio-informatics study. J Biomed Inform 41: 607-12.

George RA, Smith TD, Callaghan S, Hardman L, Pierides C, Horaitis O, Wouters MA, and Cotton RG (2008): General mutation databases: analysis and review. J Med Genet 45: 65-70.

Giardine B, Riemer C, Hefferon T, Thomas D, Hsu F, Zielenski J, Sang Y, Elnitski L, Cutting G, Trumbower H, Kern A, Kuhn R, Patrinos GP, Hughes J, Higgs D, Chui D, Scriver C, Phommarinh M, Patnaik SK, Blumenfeld O, Gottlieb B, Vihinen M, Väliaho J, Kent J, Miller W, and Hardison RC (2007): PhenCode: connecting ENCODE data with mutations and phenotype. Hum Mutat 28: 554-62.

Gillis S, Furie BC, and Furie B (1997): Interactions of neutrophils and coagulation proteins. Semin Hematol 34: 336-42.

Glaser F, Pupko T, Paz I, Bell RE, Bechor-Shental D, Martz E, and Ben-Tal N (2003): ConSurf: identification of functional regions in proteins by surface-mapping of phylogenetic information. Bioinformatics 19: 163-4.

Goeteyn M, Geerts ML, Kint A, and De Weert J (1994): The Bazex-Dupre-Christol syndrome. Arch Dermatol 130: 337-42.

Goldgar DE, Easton DF, Deffenbaugh AM, Monteiro AN, Tavtigian SV, and Couch FJ (2004): Integrated evaluation of DNA sequence variants of unknown clinical significance: application to BRCA1 and BRCA2. Am J Hum Genet 75: 535-44.

Gorlin RJ (1987) Nevoid basal-cell carcinoma syndrome. Medicine (Baltimore) 66: 98-113.

Gribenko AV, Patel MM, Liu J, McCallum SA, Wang C, and Makhatadze GI (2009): Rational stabilization of enzymes by computational redesign of surface charge-charge interactions. Proc Natl Acad Sci U S A 106: 2601-6.

Grimsley GR, Shaw KL, Fee LR, Alston RW, Huyghues-Despointes BM, Thurlkill RL, Scholtz JM, and Pace CN (1999): Increasing protein stability by altering long-range coulombic interactions. Protein Sci 8: 1843-9.

Gromiha MM, Pujadas G, Magyar C, Selvaraj S, and Simon I (2004): Locating the stabilizing residues in α/β barrel proteins based on hydrophobicity, long-range interactions, and sequence conservation. Proteins 55: 316-29.

Guijarro JI, Sunde M, Jones JA, Campbell ID, and Dobson CM (1998): Amyloid fibril formation by an SH3 domain. Proc Natl Acad Sci U S A 95: 4224-8.

Hämmerle MM, Aleksandrov AA, Chang XB, and Riordan JR (2000): A novel CFTR disease-associated mutation causes addition of an extra N-linked oligosaccharide. Glycoconj J 17: 807-13.

Hamosh A, Scott AF, Amberger JS, Bocchini CA, and McKusick VA (2005): Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Res 33: D514-7.

Han A, Kim WY, and Park SM (2007): SNP2NMD: a database of human single nucleotide polymorphisms causing nonsense-mediated mRNA decay. Bioinformatics 23: 397-9.

Harris DA and True HL (2006): New insights into prion structure and toxicity. Neuron 50: 353-7.

Hasle H, Aricò M, Basso G, Biondi A, Cantù Rajnoldi A, Creutzig U, Fenu S, Fonatsch C, Haas OA, Harbott J, Kardos G, Kerndrup G, Mann G, Niemeyer CM, Ptoszkova H, Ritter J, Slater R, Starý J, Stollmann-Gibbels B, Testi AM, van Wering ER, and Zimmermann M (1999): Myelodysplastic syndrome, juvenile myelomonocytic leukemia, and acute myeloid leukemia associated with complete or partial monosomy 7. European Working Group on MDS in Childhood (EWOG-MDS). Leukemia 13: 376-85.

Hirakawa M (2002): HOWDY: an integrated database system for human genome research. Nucleic Acids Res 30: 152-7.

Holbrook JA, Neu-Yilik G, Hentze MW, and Kulozik AE (2004): Nonsense-mediated decay approaches the clinic. Nat Genet 36: 801-8.

Hon LS, Zhang Y, Kaminker JS, and Zhang Z (2009): Computational prediction of the functional effects of amino acid substitutions in signal peptides using a model-based approach. Hum Mutat 30: 99-106.

Honig B and Nicholls A (1995): Classical electrostatics in biology and chemistry. Science 268: 1144-9.

Horaitis O, Talbot CC, Jr., Phommarinh M, Phillips KM, and Cotton RG (2007): A database of locus-specific databases. Nat Genet 39: 425.

Horovitz A, Serrano L, Avron B, Bycroft M, and Fersht AR (1990): Strength and co-operativity of contributions of surface salt bridges to protein stability. J Mol Biol 216: 1031-44.

Horwitz M, Benson KF, Person RE, Aprikyan AG, and Dale DC (1999): Mutations in ELA2, encoding neutrophil elastase, define a 21-day biological clock in cyclic haematopoiesis. Nat Genet 23: 433-6.

Hu Z, Ma B, Wolfson H, and Nussinov R (2000): Conservation of polar residues as hot spots at protein interfaces. Proteins 39: 331-42.

The International HapMap Consortium (2003): The International HapMap Project. Nature 426: 789-96.

The International Human Genome Sequencing Consortium (2001): Initial sequencing and analysis of the human genome. Nature 409: 860-921.

Ito M, Oiso Y, Murase T, Kondo K, Saito H, Chinzei T, Racchi M, and Lively MO (1993): Possible involvement of inefficient cleavage of preprovasopressin by signal peptidase as a cause for familial central diabetes insipidus. J Clin Invest 91: 2565-71.

Janin J and Chothia C (1990): The structure of protein-protein recognition sites. J Biol Chem 265: 16027-30.

Janin J, Miller S, and Chothia C (1988): Surface, subunit interfaces and interior of oligomeric proteins. J Mol Biol 204: 155-64.

Jones S and Thornton JM (1996): Principles of protein-protein interactions. Proc Natl Acad Sci U S A 93: 13-20.

Karaplis AC, Lim SK, Baba H, Arnold A, and Kronenberg HM (1995): Inefficient membrane targeting, translocation, and proteolytic processing by signal peptidase of a mutant preproparathyroid hormone protein. J Biol Chem 270: 1629-35.

Karchin R (2009): Next generation tools for the annotation of human SNPs. Brief Bioinform 10: 35-52.

Karchin R, Monteiro AN, Tavtigian SV, Carvalho MA, and Sali A (2007): Functional impact of missense variants in BRCA1 predicted by supervised learning. PLoS Comput Biol 3: e26.

Kasakov AS, Bukach OV, Seit-Nebi AS, Marston SB, and Gusev NB (2007): Effect of mutations in the β5-β7 loop on the structure and properties of human small heat shock protein HSP22 (HspB8, H11). FEBS J 274: 5628-42.

Keage HA, Carare RO, Friedland RP, Ince PG, Love S, Nicoll JA, Wharton SB, Weller RO, and Brayne C (2009): Population studies of sporadic cerebral amyloid angiopathy and dementia: a systematic review. BMC Neurol 9: 3.

Khan S and Vihinen M (2010): Performance of protein stability predictors. Hum Mutat, in press.

Khan S and Vihinen, M (2009): Evaluation of accuracy and applicability of protein models: retrospective analysis of biological and biomedical predicitons. In Silico Biology 9: 0025.

Khan S and Vihinen M (2007): Spectrum of disease-causing mutations in protein secondary structures. BMC Struct Biol 7: 56.

Khemtemourian L, Killian JA, Hoppener JW, and Engel MF (2008): Recent insights in islet amyloid polypeptide-induced membrane disruption and its role in β-cell death in type 2 diabetes mellitus. Exp Diabetes Res 2008: 421287.

Kitchens CS and Alexander JA (1983): Partial deficiency of coagulation factor XI as a newly recognized feature of Noonan syndrome. J Pediatr 102: 224-7.

Kofoed EM, Hwa V, Little B, Woods KA, Buckway CK, Tsubaki J, Pratt KL, Bezrodnik L, Jasper H, Tepper A, Heinrich JJ, and Rosenfeld RG (2003): Growth hormone insensitivity associated with a STAT5b mutation. N Engl J Med 349: 1139-47.

Kratz CP, Niemeyer CM, Castleberry RP, Cetin M, Bergstrasser E, Emanuel PD, Hasle H, Kardos G, Klein C, Kojima S, Stary J, Trebo M, Zecca M, Gelb BD, Tartaglia M, and Loh ML (2005): The mutational spectrum of PTPN11 in juvenile myelomonocytic leukemia and Noonan syndrome/myeloproliferative disease. Blood 106: 2183-5.

Krawczak M, Ball EV, and Cooper DN (1998): Neighboring-nucleotide effects on the rates of germ-line single-base-pair substitution in human genes. Am J Hum Genet 63: 474-88.

Krawczak M, Ball EV, Fenton I, Stenson PD, Abeysinghe S, Thomas N, and Cooper DN (2000): Human gene mutation database-a biomedical information and research resource. Hum Mutat 15: 45-51.

Krishnan VG and Westhead DR (2003): A comparative study of machine-learning methods to predict the effects of single nucleotide polymorphisms on protein function. Bioinformatics 19: 2199-209.

Kroczek RA, Graf D, Brugnoni D, Giliani S, Korthuer U, Ugazio A, Senger G, Mages HW, Villa A, and Notarangelo LD (1994): Defective expression of CD40 ligand on T cells causes "X-linked immunodeficiency with hyper-IgM (HIGM1)". Immunol Rev 138: 39-59.

Kudla G, Murray AW, Tollervey D, and Plotkin JB (2009): Coding-sequence determinants of gene expression in Escherichia coli. Science 324: 255-8.

Lappalainen I, Thusberg J, Shen B, and Vihinen M (2008): Genome wide analysis of pathogenic SH2 domain mutations. Proteins 72: 779-92.

Laurie AT and Jackson RM (2005): Q-SiteFinder: an energy-based method for the prediction of protein-ligand binding sites. Bioinformatics 21: 1908-16.

Laurila K and Vihinen M (2009): Prediction of disease-related mutations affecting protein localization. BMC Genomics 10: 122.

Lavergne JM, De Paillette L, Bahnak BR, Ribba AS, Fressinaud E, Meyer D, and Pietu G (1992): Defects in type IIA von Willebrand disease: a cysteine 509 to arginine substitution in the mature von Willebrand factor disrupts a disulphide loop involved in the interaction with platelet glycoprotein Ib-IX. Br J Haematol 82: 66-72.

Lee HJ, Srinivasan D, Coomber D, Lane DP, and Verma CS (2007): Modulation of the p53-MDM2 interaction by phosphorylation of Thr18: a computational study. Cell Cycle 6: 2604-11.

Lehrer RI, Ganz T, Selsted ME, Babior BM, and Curnutte JT (1988): Neutrophils and host defense. Ann Intern Med 109: 127-42.

Lemire EG (2002): Noonan syndrome or new autosomal dominant condition with coarctation of the aorta, hypertrophic cardiomyopathy, and minor anomalies. Am J Med Genet 113: 286-90.

Levy J, Espanol-Boren T, Thomas C, Fischer A, Tovo P, Bordigoni P, Resnick I, Fasth A, Baer M, Gomez L, Sanders EA, Tabone MD, Plantaz D, Etzioni A, Monafo V, Abinun M, Hammarstrom L, Abrahamsen T, Jones A, Finn A, Klemola T, DeVries E, Sanal O, Peitsch MC, Notarangelo LD (1997): Clinical spectrum of X-linked hyper-IgM syndrome. J Pediatr 131: 47-54.

Li, B., V.G. Krishnan, M.E. Mort, F. Xin, K.K. Kamati, D.N. Cooper, S.D. Mooney, and Radivojac P (2009): Automated inference of molecular mechanisms of disease from amino acid substitutions. Bioinformatics 25: 2744-50.

Lichtarge O, Bourne HR, and Cohen FE (1996): An evolutionary trace method defines binding surfaces common to protein families. J Mol Biol 257: 342-58.

Lim YP (2005): Mining the tumor phosphoproteome for cancer markers. Clin Cancer Res 11: 3163-9.

Limal JM, Parfait B, Cabrol S, Bonnet D, Leheup B, Lyonnet S, Vidaud M, and Le Bouc Y (2006): Noonan syndrome: relationships between genotype, growth, and growth factors. J Clin Endocrinol Metab 91: 300-6.

Linding R, Jensen LJ, Diella F, Bork P, Gibson TJ, and Russell RB (2003a): Protein disorder prediction: implications for structural proteomics. Structure 11: 1453-9.

Linding R, Russell RB, Neduva V, and Gibson TJ (2003b): GlobPlot: Exploring protein sequences for globularity and disorder. Nucleic Acids Res 31: 3701-8.

Lindvall JM, Blomberg KE, Väliaho J, Vargas L, Heinonen JE, Berglof A, Mohamed AJ, Nore BF, Vihinen M, and Smith CIE (2005): Bruton's tyrosine kinase: cell biology, sequence conservation, mutation spectrum, siRNA modifications, and expression profiling. Immunol Rev 203: 200-15.

Liu QR, Drgon T, Johnson C, Walther D, Hess J, and Uhl GR (2006): Addiction molecular genetics: 639,401 SNP whole genome association identifies many "cell adhesion" genes. Am J Med Genet B Neuropsychiatr Genet 141B: 918-25.

Liu R, Baase WA, and Matthews BW (2000): The introduction of strain and its effects on the structure and stability of T4 lysozyme. J Mol Biol 295: 127-45.

Loladze VV, Ermolenko DN, and Makhatadze GI (2002): Thermodynamic consequences of burial of polar and non-polar amino acid residues in the protein interior. J Mol Biol 320: 343-57.

Lovell SC, Davis IW, Arendall WB, 3rd, de Bakker PI, Word JM, Prisant MG, Richardson JS, and Richardson DC (2003): Structure validation by Cα geometry: φ, ψ and Cβ deviation. Proteins 50: 437-50.

Lovell SC, Word JM, Richardson JS, and Richardson DC (2000): The penultimate rotamer library. Proteins 40: 389-408.

Luheshi LM, Tartaglia GG, Brorsson AC, Pawar AP, Watson IE, Chiti F, Vendruscolo M, Lomas DA, Dobson CM, and Crowther DC (2007):

Systematic in vivo analysis of the intrinsic determinants of amyloid β pathogenicity. PLoS Biol 5: e290.

Ma B, Elkayam T, Wolfson H, and Nussinov R (2003): Protein-protein interactions: structurally conserved residues distinguish between binding sites and exposed protein surfaces. Proc Natl Acad Sci U S A 100: 5772-7.

MacArthur MW and Thornton JM (1991): Influence of proline residues on protein conformation. J Mol Biol 218: 397-412.

Machida K and Mayer BJ (2005): The SH2 domain: versatile signaling module and pharmaceutical target. Biochim Biophys Acta 1747: 1-25.

Matthews BW (1993): Structural and genetic analysis of protein stability. Annu Rev Biochem 62: 139-60.

Matthews BW (1995): Studies on protein stability with T4 lysozyme. Adv Protein Chem 46: 249-78.

Mészáros B, Tompa P, Simon I, amd Dosztányi Z (2007): Molecular principles of the interactions of disordered proteins. J Mol Biol 372: 549-61.

Mikkola H, Yee VC, Syrjälä M, Seitz R, Egbring R, Petrini P, Ljung R, Ingerslev J, Teller DC, Peltonen L, and Palotie A (1996): Four novel mutations in deficiency of coagulation factor XIII: consequences to expression and structure of the A-subunit. Blood 87: 141-51.

Miller MP and Kumar S (2001): Understanding human disease mutations through the use of interspecific genetic variation. Hum Mol Genet 10: 2319-28.

Minoshima S, Mitsuyama S, Ohtsubo M, Kawamura T, Ito S, Shibamoto S, Ito F, and Shimizu N (2001): The KMDB/MutationView: a mutation database for human disease genes. Nucleic Acids Res 29: 327-8.

Mirkovic N, Marti-Renom MA, Weber BL, Sali A, and Monteiro AN (2004): Structure-based assessment of missense mutations in human BRCA1: implications for breast and ovarian cancer predisposition. Cancer Res 64: 3790-7.

Mooney S (2005): Bioinformatics approaches and resources for single nucleotide polymorphism functional analysis. Brief Bioinform 6: 44-56.

Mooney SD and Altman RB (2003): MutDB: annotating human variation with functionally relevant data. Bioinformatics 19: 1858-60.

Mooney SD and Klein TE (2002): The functional importance of disease-associated mutation. BMC Bioinformatics 3: 24.

Moretti S, Armougom F, Wallace IM, Higgins DG, Jongeneel CV, and 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: W645-8.

Muñoz V and Serrano L (1997): Development of the multiple sequence approximation within the AGADIR model of α-helix formation: comparison with Zimm-Bragg and Lifson-Roig formalisms. Biopolymers 41: 495-509.

Ng PC and Henikoff S (2001): Predicting deleterious amino acid substitutions. Genome Res 11: 863-74.

Ng PC and Henikoff S (2006): Predicting the effects of amino acid substitutions on protein function. Annu Rev Genomics Hum Genet 7: 61-80.

Nichols KE, Harkin DP, Levitz S, Krainer M, Kolquist KA, Genovese C, Bernard A, Ferguson M, Zuo L, Snyder E, Buckler AJ, Wise C, Ashley J, Lovett M, Valentine MB, Look AT, Gerald W, Housman DE, and Haber DA (1998): Inactivating mutations in an SH2 domain-encoding gene in X-linked lymphoproliferative syndrome. Proc Natl Acad Sci U S A 95: 13765-70.

Noonan JA (1968): Hypertelorism with Turner phenotype. A new syndrome with associated congenital heart disease. Am J Dis Child 116: 373-80.

Notarangelo LD, Duse M, and Ugazio AG (1992): Immunodeficiency with hyper-IgM (HIM). Immunodefic Rev 3: 101-21

Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, and Thornton JM (1997): CATH--a hierarchic classification of protein domain structures. Structure 5: 1093-108.

O'Sullivan O, Suhre K, Abergel C, Higgins DG, and Notredame C (2004): 3DCoffee: combining protein sequences and structures within multiple sequence alignments. J Mol Biol 340: 385-95.

Otzen DE, Rheinnecker M, and Fersht AR (1995): Structural factors contributing to the hydrophobic effect: the partly exposed hydrophobic minicore in chymotrypsin inhibitor 2. Biochemistry 34: 13051-8.

Pace CN (1990): Conformational stability of globular proteins. Trends Biochem Sci 15: 14-7.

Pajunen M, Turakainen H, Poussu E, Peränen J, Vihinen M, and 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.

Pakula AA and Sauer RT (1989): Genetic analysis of protein stability and function. Annu Rev Genet 23: 289-310.

Panchenko AR, Kondrashov F, and Bryant S (2004): Prediction of functional sites by analysis of sequence and structure conservation. Protein Sci 13: 884-92.

Pauling L, Corey RB, and Branson HR (1951): The structure of proteins; two hydrogen-bonded helical configurations of the polypeptide chain. Proc Natl Acad Sci U S A 37: 205-11.

Pei J and Grishin NV (2001): AL2CO: calculation of positional conservation in a protein sequence alignment. Bioinformatics 17: 700-12.

Pidasheva S, Canaff L, Simonds WF, Marx SJ, and Hendy GN (2005): Impaired cotranslational processing of the calcium-sensing receptor due to signal peptide missense mutations in familial hypocalciuric hypercalcemia. Hum Mol Genet 14: 1679-90.

Piirilä H, Väliaho J, and Vihinen M (2006) Immunodeficiency mutation databases (IDbases). Hum Mutat 27: 1200-8.

Porter CT, Bartlett GJ, and Thornton JM (2004): The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data. Nucleic Acids Res 32: D129-33.

Poussu E, Vihinen M, Paulin L, and Savilahti H (2004): Probing the α-complementing domain of E. coli β-galactosidase with use of an insertional pentapeptide mutagenesis strategy based on Mu in vitro DNA transposition. Proteins 54: 681-92.

Purtilo DT (1981): X-linked lymphoproliferative syndrome. An immunodeficiency disorder with acquired agammaglobulinemia, fatal infectious mononucleosis, or malignant lymphoma. Arch Pathol Lab Med 105: 119-21.

Qu CK, Shi ZQ, Shen R, Tsai FY, Orkin SH, and Feng GS (1997): A deletion mutation in the SH2-N domain of Shp-2 severely suppresses hematopoietic cell development. Mol Cell Biol 17: 5499-507.

Qu CK, Yu WM, Azzarelli B, Cooper S, Broxmeyer HE, and Feng GS (1998): Biased suppression of hematopoiesis and multiple developmental defects in chimeric mice containing Shp-2 mutant cells. Mol Cell Biol 18: 6075-82.

Racchi M, Watzke HH, High KA, and Lively MO (1993): Human coagulation factor X deficiency caused by a mutant signal peptide that blocks cleavage by signal peptidase but not targeting and translocation to the endoplasmic reticulum. J Biol Chem 268: 5735-40.

Radivojac P, Baenziger PH, Kann MG, Mort ME, Hahn MW, and Mooney SD (2008): Gain and loss of phosphorylation sites in human cancer. Bioinformatics 24: i241-7.

Rajasekaran R, Doss GP, Sudandiradoss C, Ramanathan K, Rituraj P, and Sethumadhavan R (2008): Computational and structural investigation of deleterious functional SNPs in breast cancer BRCA2 gene. Sheng Wu Gong Cheng Xue Bao 24: 851-6.

Rajasekaran R, Sudandiradoss C, Doss CG, and Sethumadhavan R (2007): Identification and in silico analysis of functional SNPs of the BRCA1 gene. Genomics 90: 447-52.

Ramana CV, Chatterjee-Kishore M, Nguyen H, and Stark GR (2000): Complex roles of Stat1 in regulating gene expression. Oncogene 19: 2619-27.

Ramensky V, Bork P, and Sunyaev S (2002): Human non-synonymous SNPs: server and survey. Nucleic Acids Res 30: 3894-900.

Rawlings DJ and Witte ON (1994): Bruton's tyrosine kinase is a key regulator in B-cell development. Immunol Rev 138: 105-19.

Riikonen P and Vihinen M (1999): MUTbase: maintenance and analysis of distributed mutation databases. Bioinformatics 15: 852-9.

Ring HZ, Kwok PY, and Cotton RG (2006): Human Variome Project: an international collaboration to catalogue human genetic variation. Pharmacogenomics 7: 969-72.

Robinson PA (2008): Protein stability and aggregation in Parkinson's disease. Biochem J 413: 1-13.

Romero, Obradovic, and Dunker K (1997): Sequence Data Analysis for Long Disordered Regions Prediction in the Calcineurin Family. Genome Inform Ser Workshop Genome Inform 8: 110-124.

Sankararaman S and Sjolander K (2008): INTREPID--INformation-theoretic TREe traversal for Protein functional site IDentification. Bioinformatics 24: 2445-52.

Saunders CT and Baker D (2002): Evaluation of structural and evolutionary contributions to deleterious mutation prediction. J Mol Biol 322: 891-901.

Saxton TM, Ciruna BG, Holmyard D, Kulkarni S, Harpal K, Rossant J, and Pawson T (2000): The SH2 tyrosine phosphatase shp2 is required for mammalian limb development. Nat Genet 24: 420-3.

Saxton TM, Henkemeyer M, Gasca S, Shen R, Rossi DJ, Shalaby F, Feng GS, and Pawson T (1997): Abnormal mesoderm patterning in mouse embryos mutant for the SH2 tyrosine phosphatase Shp-2. EMBO J 16: 2352-64.

Sayos J, Wu C, Morra M, Wang N, Zhang X, Allen D, van Schaik S, Notarangelo L, Geha R, Roncarolo MG, Oettgen H, De Vries JE, Aversa G, and Terhorst C (1998): The X-linked lymphoproliferative-disease gene product SAP regulates signals induced through the co-receptor SLAM. Nature 395: 462-9.

Schimmel PR and Flory PJ (1968): Conformational energies and configurational statistics of copolypeptides containing L-proline. J Mol Biol 34: 105-20.

Schlessinger J and Lemmon MA (2003): SH2 and PTB domains in tyrosine kinase signaling. Sci STKE 191: RE12.

Schneider G and Fechner U (2004): Advances in the prediction of protein targeting signals. Proteomics 4: 1571-80.

Schymkowitz J, Borg J, Stricher F, Nys R, Rousseau F, and Serrano L (2005): The FoldX web server: an online force field. Nucleic Acids Res 33: W382-8.

Seeliger MW, Biesalski HK, Wissinger B, Gollnick H, Gielen S, Frank J, Beck S, and Zrenner E (1999): Phenotype in retinol deficiency due to a hereditary defect in retinol binding protein synthesis. Invest Ophthalmol Vis Sci 40: 3-11.

Selkoe DJ (1996): Amyloid β-protein and the genetics of Alzheimer's disease. J Biol Chem 271: 18295-8.

Seppen J, Steenken E, Lindhout D, Bosma PJ, and Elferink RP (1996) A mutation which disrupts the hydrophobic core of the signal peptide of bilirubin UDP-glucuronosyltransferase, an endoplasmic reticulum membrane protein, causes Crigler-Najjar type II. FEBS Lett 390: 294-8.

Sharma A, Chavali S, Mahajan A, Tabassum R, Banerjee V, Tandon N, and Bharadwaj D (2005): Genetic association, post-translational modification, and protein-protein interactions in Type 2 diabetes mellitus. Mol Cell Proteomics 4: 1029-37.

Shen B and Vihinen M (2003): RankViaContact: ranking and visualization of amino acid contacts. Bioinformatics 19: 2161-2.

Shen B and 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: 267-76.

Shen J, Deininger PL, and Zhao H (2006): Applications of computational algorithm tools to identify functional SNPs in cytokine genes. Cytokine 35: 62-6.

Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, and Sirotkin K (2001): dbSNP: the NCBI database of genetic variation. Nucleic Acids Res 29: 308-11.

Shirley BA, Stanssens P, Hahn U, and Pace CN (1992): Contribution of hydrogen bonding to the conformational stability of ribonuclease T1. Biochemistry 31: 725-32.

Shortle D, Stites WE, and Meeker AK (1990): Contributions of the large hydrophobic amino acids to the stability of staphylococcal nuclease. Biochemistry 29: 8033-41.

Snyder JA, Weston A, Tinkle SS, and Demchuk E (2003) Electrostatic potential on human leukocyte antigen: implications for putative mechanism of chronic beryllium disease. Environ Health Perspect 111: 1827-34.

Sobolev V, Sorokine A, Prilusky J, Abola EE, and Edelman M (1999): Automated analysis of interatomic contacts in proteins. Bioinformatics 15: 327-32.

Sonnhammer EL, Eddy SR, and Durbin R (1997): Pfam: a comprehensive database of protein domain families based on seed alignments. Proteins 28: 405-20.

Stenson PD, Ball EV, Mort M, Phillips AD, Shiel JA, Thomas NS, Abeysinghe S, Krawzak M, and Cooper DN (2003): Human Gene Mutation Database (HGMD): 2003 update. Hum Mutat 21: 577-81.

Stenson PD, Mort M, Ball EV, Howells K, Phillips AD, Thomas NS, and Cooper DN (2009): The Human Gene Mutation Database: 2008 update. Genome Med 1: 13.

Steward RE, MacArthur MW, Laskowski RA, and Thornton JM (2003): Molecular basis of inherited diseases: a structural perspective. Trends Genet 19: 505-13.

Stone EA and Sidow A (2005): Physicochemical constraint violation by missense substitutions mediates impairment of protein function and disease severity. Genome Res 15: 978-86.

Strickler SS, Gribenko AV, Keiffer TR, Tomlinson J, Reihle T, Loladze VV, and Makhatadze GI (2006): Protein stability and surface electrostatics: a charged relationship. Biochemistry 45: 2761-6.

Sunyaev S, Ramensky V, and 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 W 3rd, Kondrashov AS, and Bork P (2001): Prediction of deleterious human alleles. Hum Mol Genet 10: 591-7.

Tanford C (1978): The hydrophobic effect and the organization of living matter. Science 200: 1012-8.

Tang H, Wyckoff GJ, Lu J, and Wu CI (2004): A universal evolutionary index for amino acid changes. Mol Biol Evol 21: 1548-56.

Tang TL, Freeman RM, Jr., O'Reilly AM, Neel BG, and Sokol SY (1995): The SH2-containing protein-tyrosine phosphatase SH-PTP2 is required upstream of MAP kinase for early Xenopus development. Cell 80: 473-83.

Tartaglia M, Niemeyer CM, Fragale A, Song X, Buechner J, Jung A, Hahlen K, Hasle H, Licht JD, and Gelb BD (2003): Somatic mutations in PTPN11 in juvenile myelomonocytic leukemia, myelodysplastic syndromes and acute myeloid leukemia. Nat Genet 34: 148-50.

Terp BN, Cooper DN, Christensen IT, Jørgensen FS, Bross P, Gregersen N, and 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.

Thermann R, Neu-Yilik G, Deters A, Frede U, Wehr K, Hagemeier C, Hentze MW, and Kulozik AE (1998): Binary specification of nonsense codons by splicing and cytoplasmic translation. EMBO J 17: 3484-94.

Thomas PD, Campbell MJ, Kejariwal A, Mi H, Karlak B, Daverman R, Diemer K, Muruganujan A, and Narechania A (2003): PANTHER: a library of protein families and subfamilies indexed by function. Genome Res 13: 2129-41.

Thomas PJ, Qu BH, and Pedersen PL (1995): Defective protein folding as a basis of human disease. Trends Biochem Sci 20: 456-9.

Thompson JD, Higgins DG, and 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: 4673-80.

Thorisson GA, Lancaster O, Free RC, Hastings RK, Sarmah P, Dash D, Brahmachari SK, and Brookes AJ (2009): HGVbaseG2P: a central genetic association database. Nucleic Acids Res 37: D797-802.

Trojanowski JQ and Lee VM (1998): Aggregation of neurofilament and α-synuclein proteins in Lewy bodies: implications for the pathogenesis of Parkinson disease and Lewy body dementia. Arch Neurol 55: 151-2.

The UniProt Consortium (2010): The Universal Protein Resource (UniProt) in 2010. Nucleic Acids Res 38: D142-8.

Uversky VN, Oldfield CJ, and Dunker AK (2008): Intrinsically disordered proteins in human diseases: introducing the D2 concept. Annu Rev Biophys 37: 215-46.

van Hoof A, Frischmeyer PA, Dietz HC, and Parker R (2002) Exosome-mediated recognition and degradation of mRNAs lacking a termination codon. Science 295: 2262-4.

Ventura S, Zurdo J, Narayanan S, Parreño M, Mangues R, Reif B, Chiti F, Giannoni E, Dobson CM, Aviles FX, and Serrano L (2004): Short amino acid stretches can mediate amyloid formation in globular proteins: the Src homology 3 (SH3) case. Proc Natl Acad Sci U S A 101: 7258-63.

Vitkup D, Sander C, and Church GM (2003): The amino-acid mutational spectrum of human genetic disease. Genome Biol 4: R72.

Vogt G, Chapgier A, Yang K, Chuzhanova N, Feinberg J, Fieschi C, Boisson-Dupuis S, Alcais A, Filipe-Santos O, Bustamante J, de Beaucoudrey L, Al-Mohsen I, Al-Hajjar S, Al-Ghonaium A, Adimi P, Mirsaeidi M, Khalilzadeh S, Rosenzweig S, de la Calle Martin O, Bauer TR, Puck JM, Ochs HD, Furthner D, Engelhorn C, Belohradsky B, Mansouri D, Holland SM, Schreiber RD, Abel L, Cooper DN, Soudais C, and Casanova JL (2005): Gains of glycosylation comprise an unexpectedly large group of pathogenic mutations. Nat Genet 37: 692-700.

Vogt G, Vogt B, Chuzhanova N, Julenius K, Cooper DN, and Casanova JL (2007): Gain-of-glycosylation mutations. Curr Opin Genet Dev 17: 245-51.

Vriend G (1990): WHAT IF: a molecular modeling and drug design program. J Mol Graph 8: 52-6, 29.

Wang Z and Moult J (2001): SNPs, protein structure, and disease. Hum Mutat 17: 263-70.

Ward JJ, McGuffin LJ, Bryson K, Buxton BF, and Jones DT (2004): The DISOPRED server for the prediction of protein disorder. Bioinformatics 20: 2138-9.

Weiss SJ (1989): Tissue destruction by neutrophils. N Engl J Med 320: 365-76.

Williams RS, Chasman DI, Hau DD, Hui B, Lau AY, and Glover JN (2003): Detection of protein folding defects caused by BRCA1-BRCT truncation and missense mutations. J Biol Chem 278: 53007-16.

Williams RS and Glover JN (2003): Structural consequences of a cancer-causing BRCA1-BRCT missense mutation. J Biol Chem 278: 2630-5.

Wintroub BU, Coblyn JS, Kaempfer CE, and Austen KF (1980): Cleavage of fibrinogen by the human neutrophil neutral peptide-generating protease. Proc Natl Acad Sci U S A 77: 5448-5.

Woelfle J, Billiard J, and Rotwein P (2003): Acute control of insulin-like growth factor-I gene transcription by growth hormone through Stat5b. J Biol Chem 278: 22696-702.

Word JM, Bateman RC, Jr., Presley BK, Lovell SC, and Richardson DC (2000): Exploring steric constraints on protein mutations using MAGE/PROBE. Protein Sci 9: 2251-9.

Word JM, Lovell SC, LaBean TH, Taylor HC, Zalis ME, Presley BK, Richardson JS, and Richardson DC (1999): Visualizing and quantifying molecular goodness-of-fit: small-probe contact dots with explicit hydrogen atoms. J Mol Biol 285: 1711-33.

Wright PE and Dyson HJ (1999) Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. J Mol Biol 293: 321-31.

Xu J, Baase WA, Baldwin E, and Matthews BW (1998): The response of T4 lysozyme to large-to-small substitutions within the core and its relation to the hydrophobic effect. Protein Sci 7: 158-77.

Yamaguchi-Kabata Y, Shimada MK, Hayakawa Y, Minoshima S, Chakraborty R, Gojobori T, and Imanishi T (2008): Distribution and effects of nonsense polymorphisms in human genes. PLoS ONE 3: e3393.

Yang ZR, Thomson R, McNeil P, and Esnouf RM (2005): RONN: the bio-basis function neural network technique applied to the detection of natively disordered regions in proteins. Bioinformatics 21: 3369-76.

Yankner BA and Lu T (2009): Amyloid β-protein toxicity and the pathogenesis of Alzheimer disease. J Biol Chem 284: 4755-9.

Yip YL, Scheib H, Diemand AV, Gattiker A, Famiglietti LM, Gasteiger E, and Bairoch A (2004): The Swiss-Prot variant page and the ModSNP database: a resource for sequence and structure information on human protein variants. Hum Mutat 23: 464-70.

Yue P, Li Z, and Moult J (2005): Loss of protein structure stability as a major causative factor in monogenic disease. J Mol Biol 353: 459-73.

Yue P, Melamud E, and Moult J (2006): SNPs3D: candidate gene and SNP selection for association studies. BMC Bioinformatics 7: 166.

Zhang EE, Chapeau E, Hagihara K, and Feng GS (2004): Neuronal Shp2 tyrosine phosphatase controls energy balance and metabolism. Proc Natl Acad Sci U S A 101: 16064-9.

Zhang J, Sun X, Qian Y, LaDuca JP, and Maquat LE (1998): At least one intron is required for the nonsense-mediated decay of triosephosphate isomerase mRNA: a possible link between nuclear splicing and cytoplasmic translation. Mol Cell Biol 18: 5272-83.

Zhou H and 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: 2714-26.

Zhou HX and Shan Y (2001): Prediction of protein interaction sites from sequence profile and residue neighbor list. Proteins 44: 336-43.

Zvelebil MJ, Barton GJ, Taylor WR, and Sternberg MJ (1987): Prediction of protein secondary structure and active sites using the alignment of homologous sequences. J Mol Biol 195: 957-61.

RESEARCH ARTICLE

Bioinformatic Analysis of Protein Structure–Function Relationships: Case Study of Leukocyte Elastase (*ELA2*) Missense Mutations

Janita Thusberg[1] and Mauno Vihinen[1,2]*

[1]Institute of Medical Technology, University of Tampere, Finland; [2]Research Unit, Tampere University Hospital, Tampere, Finland

For the Immunogenetics Special Issue

Cyclic and congenital neutropenia are caused by mutations in the human neutrophil elastase (HNE) gene (*ELA2*), leading to an immunodeficiency characterized by decreased or oscillating levels of neutrophils in the blood. The HNE mutations presumably cause loss of enzyme activity, consequently leading to compromised immune system function. To understand the structural basis for the disease, we implemented methods from bioinformatics to analyze all the known HNE missense mutations at both the sequence and structural level. Our results demonstrate that the 32 different mutations have diverse effects on HNE structure and function, affecting structural disorder and aggregation tendencies, stability maintaining contacts, and electrostatic properties. A large proportion of the mutations are located at conserved amino acids, which are usually essential in determining protein structure and function. The majority of the disease-causing HNE missense mutations lead to major structural changes and loss of stability in the protein. A few mutations also affect functional residues, leading into decreased catalytic activity or altered ligand binding. Our analysis reveals the putative effects of all known missense mutations in HNE, thus allowing the structural basis of cyclic and congenital neutropenia to be elucidated. We have employed and analyzed a set of some 30 different methods for predicting the effects of amino acid substitutions. We present results and experience from the analysis of the applicability of these methods in the analysis of numerous genes, proteins, and diseases to reveal protein structure–function relationships and disease genotype–phenotype correlations. Hum Mutat 27(12), 1230–1243, 2006. © 2006 Wiley-Liss, Inc.

KEY WORDS: immunology; immunogenetics; human neutrophil elastase; *ELA2*; neutropenia; structure–function relationships; structural basis of diseases

INTRODUCTION

Human cyclic neutropenia (MIM# 162800) is an autosomal dominant disease in which blood cell production from the bone marrow oscillates with 21-day periodicity [Haurie et al., 1998; Palmer et al., 1996]. During the intervals of neutropenia, affected individuals are at risk for opportunistic infections [Dale and Hammond, 1988]. Both cyclic and congenital neutropenia (MIM# 202700) are caused by mutations in the *ELA2* gene (MIM# 130130), which encodes the human neutrophil elastase (HNE) protein (SWISSPROT P08246) [Horwitz et al., 1999; Dale et al., 2000].

The azurophil granules in neutrophils, also called primary granules, contain serine proteases HNE, cathepsin G (CG), and proteinase 3 (PR3) in high concentrations [Wiedow et al., 1996]. The granules fuse with neutrophil phagolysosomes after ingestion of foreign materials as a part of the phagocytic process [Dewald et al., 1975]. The proteases are responsible for the degradation of the internalized objects [Lehrer et al., 1988; Boxer and Morganroth, 1987]. The HNE also acts extracellularly, particularly at sites of inflammation [Weiss, 1989]. The human neutrophil elastase degrades not only elastin, but also various other tissue proteins such as collagens, proteoglycans, and plasma factors like fibrinogen, fibrin, and antithrombin III [Wintroub et al., 1980; Bach-Gansmo et al., 1996; Gillis et al., 1997], and bacteria virulence proteins [Belaaouaj et al., 2000; Weinrauch et al., 2002]. A recently described novel mechanism explains how neutrophils release granule proteins and chromatin, which together form extracellular fibers to bind and kill bacteria [Brinkmann et al., 2004]. Mutations in the HNE presumably cause loss or reduction of the enzyme activity, consequently leading into compromised immune system function.

HNE, encoded by the *ELA2* gene, is a member of the trypsin family. It is a 30-kDa serine protease predominantly expressed in neutrophil granules [Bieth, 1998]. The 267-residue protein is posttranslationally processed at both termini. The N-terminal signal peptide (residues 1–27), as well as the C-terminal stretch (amino acids 248–267), is eliminated by proteolysis during processing the mature enzyme. The propeptide, amino acids

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

Grant sponsor: Tampere Graduate School in Biomedicine and Biotechnology (TGSBB); Grant sponsor: Medical Research Fund of the Tampere University Hospital, European Union.

DOI 10.1002/humu.20407

Published online 19 September 2006 in Wiley InterScience (www.interscience.wiley.com).

© 2006 WILEY-LISS, INC.

28–29, is cleaved off by the protease cathepsin C, in a step corresponding to zymogen activation [Adkinson et al., 2002; Salvesen and Enghild, 1990]. The remaining part of the protein, the activated enzyme, consists of the leukocyte elastase domain that is homologous to trypsin (amino acids 30–267). HNE folds into two structurally similar, antiparallel β-barrel domains, and the catalytic residues are localized in the crevice between the domains. Across the crevice is the substrate-binding site, which includes parts of both domains [Bode et al., 1986]. The crystal structure of the HNE trypsin homology domain has been determined to an 1.8 Å resolution [Bode et al., 1986].

There is great diversity in the structures and properties of serine proteases owing to the heterogeneity of their origins [Barrett and Rawlings, 1995]. Common for all the trypsin-like serine proteases is the catalytic triad (H70, D117, S202; HNE numbering), which is arranged in a highly conserved three-dimensional structure, providing the universal mechanism by which serine proteases cleave their substrates [Kraut, 1977; Neurath, 1986]. Regions not directly involved in the catalytic activity of the protease vary significantly between the members of the serine protease superfamily, reflecting their different biological functions (sequence homology 30–40%; reviewed by Bode et al. [1989]).

Our mutation registry for cyclic and congenital neutropenia, ELA2base (http://bioinf.uta.fi/ELA2base), currently lists 135 patient entries with a total of 43 different HNE mutations. Most disease causing mutations are found in exons (38), 32 of which are missense mutations located mainly in the trypsin homology domain of the protein (Table 1). We investigated the protein structural consequences of all the HNE missense mutations by applying structural and bioinformatics methods and herein we discuss bioinformatics tools available for protein structure–function studies and prediction and evaluation of the effects of missense mutations.

When exploring the effects of mutations on protein structure and function, several aspects should be taken into account. A disease phenotype can arise when an amino acid substitution results in the loss of a critical protein function or structural alterations. Even minor changes in the size or properties of the side chain can alter or prevent the function of the protein. On the other hand, protein folds are rather robust and allow insertions to numerous sites without the loss of function [Poussu et al., 2004]. The effect of alterations varies by the type of the mutation and the sequence and structure context. A mutation may also lead into gain of function effects, such as functional dysregulation or the formation of toxic aggregates [Forloni et al., 2002; Dobson, 2003; Steward et al., 2003; Sanders and Myers, 2004]. Mutations to posttranslationally modified positions also lead to diseases, including immunodeficiencies [Aghamohammadi et al., 2004; Vogt et al., 2005]. Evolutionarily conserved positions usually have a critical role in protein structure and function [Miller and Kumar, 2001; Mooney and Klein, 2002; Shen and Vihinen, 2004]. Amino acid substitutions at certain positions may lead into increased disorder in the structure or aggregation tendency, or the introduced side chains may have effects on protein scaffolding and stability through steric clashes, altered charge, or loss of critical interactions. Mutation-induced changes in electrostatic surface potentials have a wide-ranging effect on protein folding and stability and on protein–protein interactions. Knowledge of the regions essential for specificity and function of the protein is also needed, since mutations at these positions usually lead to loss of or decrease in activity. Here we have applied and evaluated the use and suitability of sequence and structure based methods for the analysis of the consequences of disease-related missense mutations.

MATERIALS AND METHODS

The amino acid sequence and missense mutations for the leukocyte elastase were obtained from our ELA2base database (http://bioinf.uta.fi/ELA2base). The database lists all known mutations for the *ELA2* gene and stores patient information, including clinical and immunological phenotype. The database has been built based on the principles applied in our other immunodeficiency mutation databases [Piirilä and Vihinen, 2006]. A total of 75 sequence homologs for the trypsin domain (residues 30–242) and sequence alignments were from the Pfam database [Bateman et al., 2004]. Alignments were visualized using MultiDisp (Riikonen, P., Horvath, T., and Vihinen, M., unpublished results) and ConSeq [Berezin et al., 2004] for illustration of conserved amino acids in the sequence. The default parameters were applied in all methods, if not otherwise stated.

The evolutionary conservation of the sequences was studied, in addition to the visualization programs, by ProCon, a program for calculating mutual information and entropy in amino acid sequences [Shen and Vihinen, 2004]. Conservation indices were calculated with the program al2co [Pei and Grishin, 2001] and the ConSurf server [Glaser et al., 2003].

Structural disorder in the protein and the effects of mutations were studied using six predictors, Disopred [Ward et al., 2004a], DisEMBL [Linding et al., 2003a], Globplot [Linding et al., 2003b], IUPred [Dosztányi et al., 2005a,b], DRIP-PRED, a web-based predictor for disordered regions in proteins (www.sbc.su.se/~maccallr/disorder) [MacCallum, 2006], and Ronn [Yang et al., 2005]. We compared the results for the wild-type sequence to those for mutant sequences.

The effects of mutations on aggregation propensities were studied by TANGO [Fernandez-Escamilla et al., 2004; Linding et al., 2004] and calculations presented by Chiti et al. [2003], for which α helical propensities were calculated with the program AGADIR [Muños and Serrano, 1994; Lacroix et al., 1998]. A script was written to implement the method of Chiti et al. [2003].

The damaging effects of point mutations were analyzed using SIFT [Ng and Henikoff, 2001], PolyPhen [Sunyaev et al., 2001; Ramensky et al., 2002], and Pmut [Ferrer-Costa et al., 2005].

The effects of mutations on protein stability were predicted by Scpred [Dosztányi et al., 1997], SCide [Dosztányi et al., 1997, 2003], Sride [Gromiha et al., 2004], PoPMuSiC [Gilis and Rooman, 2000; Kwasigroch et al., 2002], FoldX [Schymkowitz et al., 2005], and Dmutant [Zhou and Zhou, 2002].

Structural analyses were performed based on the crystal structure of the protein (PDB 1PPF). The structure was visualized and the mutations were modeled by PyMOL [DeLano, 2002]. Hydrogen atoms were added to the structures using Reduce [Word et al., 1999a]. Mutant amino acid side chain χ angles were rotated at intervals of 10° by the Autobondrot function in PROBE [Word et al., 1999b, 2000] and the best rotamers were selected for further analysis. The acceptable conformations for a mutated side chain have a total score above –1.0, allowing for small local perturbations in the structure [Lovell et al., 2000]. The created structures were verified by MolProbity [Lovell et al., 2003]. MolProbity adds all atom contacts into the structures and flips asparagine and glutamine side chains when necessary. Mutated structures were visualized by the program KiNG [Lovell et al., 2003], with which all atom contacts and clashes were analyzed.

Amino acid contact analysis for the mutant residues in the trypsin homology domain was performed with CSU [Sobolev et al., 1999], and the nature of the contacts, contact surfaces, as well as solvent accessible surfaces, were elucidated. Contact energies between amino acids in the trypsin homology domain were analyzed using RankViaContact [Shen and Vihinen, 2003]. By analyzing the wild-type protein, we could determine structurally important amino acids, which contribute to the stability of the protein, or amino acids with strong contacts that may be important for functional specificity. The analysis of changes in the contact energies for mutant structures provided hypotheses for the roles of the mutated amino acids. Electrostatic surface potentials were calculated and visualized with the PyMOL program [DeLano, 2002] using the absolute electrostatic potential in a vacuum. For web resources for the methods, see Table 2.

RESULTS

Sequence Conservation

Disease-causing mutations are typically located at conserved positions within a protein family, since these positions are usually essential for the structure and/or function of the protein [Miller and Kumar, 2001; Mooney and Klein, 2002; Shen and Vihinen, 2004]. In the case of HNE, there are several homologs for the trypsin homology domain, the signal peptide being less conserved. Our alignments are for the trypsin homology domain only. There is one invariant position in the protein family, G203 in HNE. There are no known disease-causing missense mutations at this position. However, several missense mutations affect positions that are highly conserved by the agreement of all the methods used: P42, C55, A57, C71, G85, G214, and P205 (Fig. 1). Many missense mutations occur at positions where certain amino acid physicochemical properties are conserved in the family. A total of 15 of the HNE missense mutations affect residues whose hydrophobicity is highly conserved in the protein family. These residues are mostly buried (the solvent accessible surface being 0.0–4.7 Å^2, except for position 43, where the solvent accessible surface is 25.1 Å^2). Introduction of charged and polar amino acids, as well as residues that may change the orientation of the main chain, such as prolines, likely has a destabilizing effect at these positions. A57T, I60T, C71S, and C151S/Y mutations introduce polar residues; C71R, G85E, and G214R introduce charged side chains; and L47P and L121P place prolines into positions where hydrophobicity is a conserved property. In the positions 43, 101, 82, 206, and 219 the conserved hydrophobic nature of the positions does not change, because the substituting amino acids are hydrophobic as well (Fig. 1). Several methods have been developed to predict hydropathic characteristics of proteins. However, these predictions, just as those for secondary structures, are of such low accuracy (Jääskeläinen, S., Riikonen, P., Salakoski, T., and Vihinen, M., unpublished results) that they cannot be used to analyze point mutations. There are 22 covarying residues in the protein family, five of which are mutated in cyclic and congenital neutropenia (Fig. 2). The majority of HNE missense mutations occur at the conserved positions (Table 1).

Mutations Predicted to Affect Structural Disorder and β Aggregation

The effects of mutations on structural disorder were investigated by six different methods. None of the mutations was predicted to increase disorder by all the programs used. We consider the following mutations likely to increase disorder in HNE because they were predicted to do so by at least three independent methods: L47P, I60T, R81P, V82M, L121P, A127P, C151S, and L152P.

Substitutions S46F, H53L, A127P, and G210V were predicted to increase β aggregation propensity in HNE when calculated by the method of Chiti et al. [2003], and P42L, S46F, and G210V were predicted to do so by the program TANGO (Table 1) [Fernandez-Escamilla et al., 2004; Linding et al., 2004].

Mutations Affecting Protein Structural Stability

The loss of structural stability caused by mutations was predicted by the programs Scide [Dosztányi et al., 1997, 2003] and Scpred [Dosztányi et al., 1997], which predict stability centers in proteins, as well as by Sride [Gromiha et al., 2004], which predicts stabilizing residues. The programs Dmutant [Zhou and Zhou, 2002], FoldX [Schymkowitz et al., 2005], and PoPMusiC [Gilis and Rooman, 2000; Kwasigroch et al., 2002] evaluate stability changes in proteins caused by mutations. A number of disease-causing mutations are located at the predicted stability centers, including S46F, L47P, A57T, R81P, V82M, V101M, L121P, L152P, and G214R. These mutations are likely to affect protein stability. Mutations S46F, L47P, A57T, I60T, C71R, C71S, R81P, V82M, L84P, G85E, L121P, C151S, C151Y, L152P, P205R, and R220Q were predicted to cause loss of stability by at least three independent methods that predict destabilizing effects of individual residues (Table 1; together with effects on essential interresidue contacts). A substitution at a predicted stabilizing residue was considered to destabilize structure.

Structural Mutations

The effects of mutations on protein structure and stability were studied by rotamer analysis and determination of clashing side chains. The best rotamers were used in the analyses. Most of the mutated side chains do not fit into the structure without deleterious changes to protein scaffolding, as determined both computationally by the Probe score [Word et al., 1999b, 2000] and by visual inspection. Of the 23 mutations screened, excluding mutations introducing proline side chains, as well as mutations not positioned at the structurally determined trypsin domain, only seven gave an acceptable score above –1.0, which allows for small local perturbations to take place in the structure [Lovell et al., 2000]. Mutations H53L, C55Y, I60T, and V219I also caused serious clashes with other side chains in spite of their positive Probe scores, leaving V101M, S126L, and R220Q the only mutations that do not cause significant rearrangements to the structure (Table 1). Mutated amino acids that cannot fit into the structure without clashes, lead to changes in protein scaffolding, stability, and properties of the protein. Proline is a known secondary structure breaker, and L47P, L84P, and L121P are located in the middle of β strands, whereas R81 precedes a β strand.

Mutations Causing Changes in Contacts Maintaining Stability

Amino acids located in the core of the protein, with a negligible solvent accessible surface area, typically form several hydrophobic interactions essential for the folding of the protein and for the stability of the protein structure. Mutations to such residues may cause detrimental changes to the structure-maintaining contacts. The probability of a mutation to be pathogenic has been shown to increase with a decrease in the solvent accessibility of the site [Vitkup et al., 2003]. In addition, the introduction of charged side chains into the hydrophobic core destabilizes protein structure [Chasman and Adams, 2001]. Of the 29 HNE trypsin homology domain mutations, four cause significant loss of hydrophobic interactions: L47P, I60T, C71S, and L206F. All these residues are located in the hydrophobic core of the protein, with a solvent accessible surface of 0.0 to 4.7%. Mutations in cysteines 71, 55,

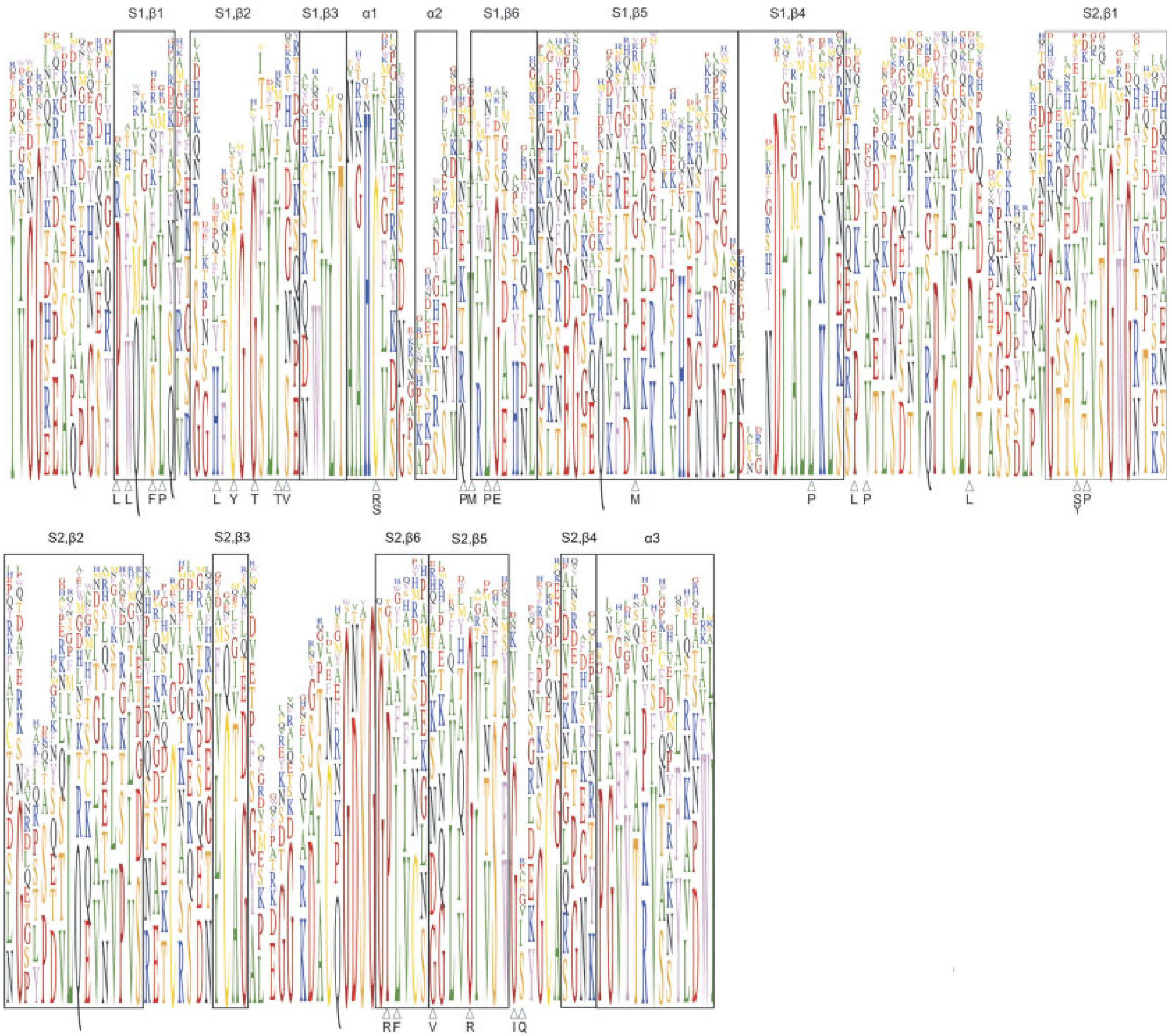


FIGURE 1. **MultiDisp visualization of the sequence alignment for the HNE trypsin homology domain and its homologues. The height of the characters indicates the frequency of the amino acids in the alignment positions, and the color of the objects reflects the chemical nature of the amino acids (physicochemically related amino acids have the same color). The secondary structures according to HNE structure are presented as boxes, and the positions of HNE missense mutations are indicated by arrowheads below the alignment, together with all mutant forms.**

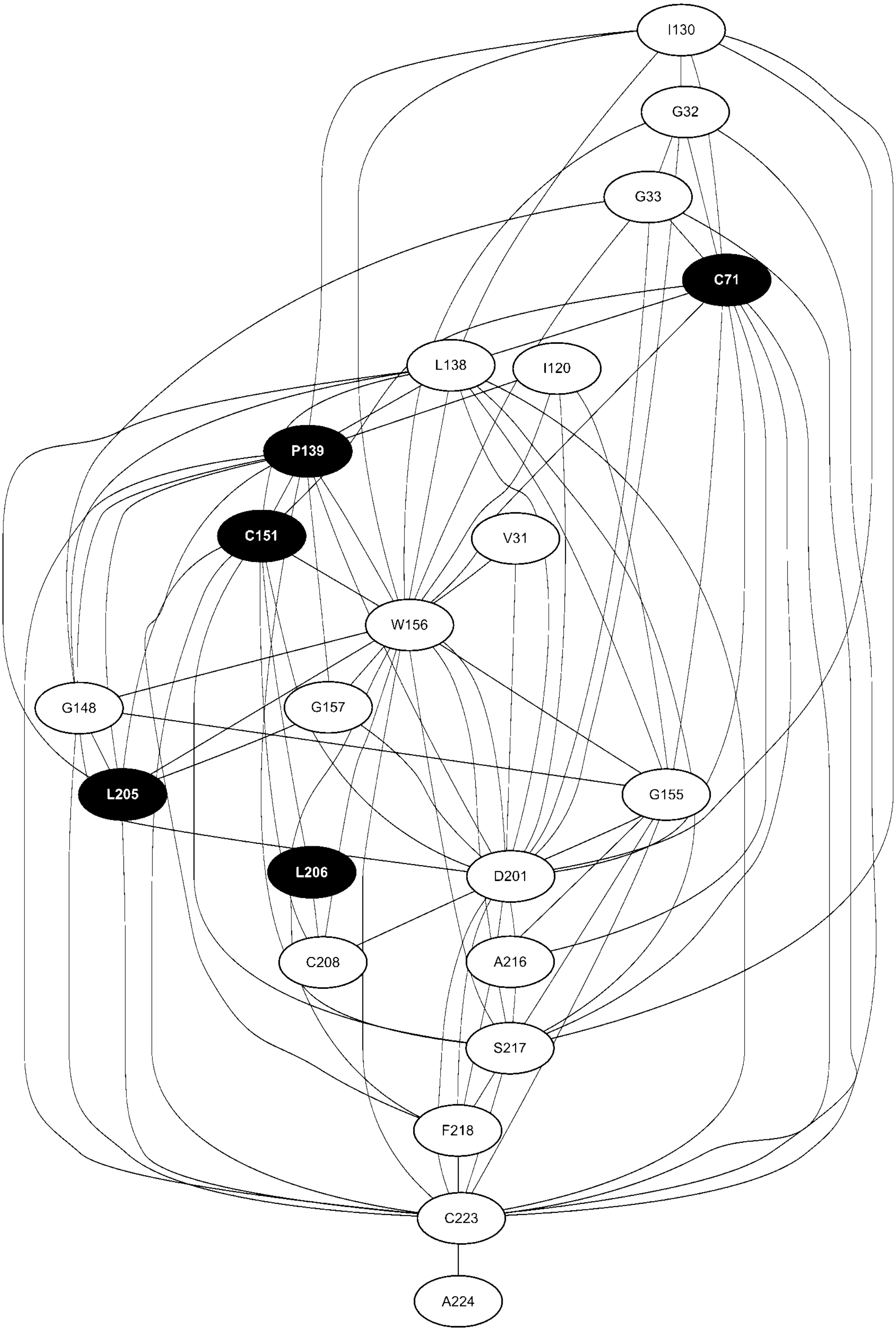


FIGURE 2. **Covarying residues in the trypsin homology protein family calculated with the program ProCon (p-value 0.001). The residues and numbers correspond to HNE. The residues in which there are missense mutations in HNE, are highlighted black.**

and 151 lead to loss of stability-maintaining disulfide bridges (Table 1).

Residues with strong contact energies are important for protein stability [Shen and Vihinen, 2003]. Mutations that affect such residues can thus be assumed to decrease stability. Of the 29 investigated mutations, seven affect residues from among the 10% most stabilizing amino acids: A57T, L121P, C151S, C151Y, L152P, L206F, and G214R.

TABLE 1. Summary of the Effects of HNE Mutations on Protein Structure and Function

Mutation	Conserved residue	Pheno-type	Electrostatic surface potential	Contacts and stability	Disorder	Conformational	Aggre-gation	Functional mutations	Other
M1V		1							Abnormal expression[a]
P42L	X	1				X	X		
F43L	X	2	X			X			
S46F		2		X		X	X		
L47P	X	1		X	X				
H53L		1		X		X	X	X	
C55Y	X	1	X			X		X	TM domain[b]; disulfide bond
A57T	X	1	X	X		X			TM domain
I60T	X	1	X	X	X	X			TM domain
A61V		1,2	X			X			
C71R	X	1		X		X		X	Disulfide bond
C71S	X	1		X		X		X	Disulfide bond
R81P		1	X	X	X				
V82M	X	2		X	X	X			
L84P		1	X	X					
G85E	X	1		X		X			
V101M	X	1		X					
L121P	X	1		X	X				
S126L		1,2							
A127P		1			X		X		
P139L	X	1,2				X			
C151S	X	1		X	X	X			Disulfide bond
C151Y	X	1		X		X			Disulfide bond
L152P		1		X	X				
P205R	X	1		X		X			TM domain
L206F	X	2		X		X			TM domain
G210V		1				X	X		TM domain
G214R	X	1		X		X			
V219I	X	2		X		X		X	TM domain
R220Q		2	X	X				X	
P257L		1							Protein trafficking[b]
P262L		2							Protein trafficking[b]

[a] Bellanné-Chantelot et al. [2004].
[b] Benson et al. [2003].
Phenotype 1, congenital neutropenia; Phenotype 2, cyclic neutropenia; TM, transmembrane domain.

Mutations Affecting the Electrostatic Surface Potential of HNE

Eight of the 29 missense mutations in the trypsin homology domain cause changes in the electrostatic surface potential of the molecule: A57T, A61V, C55Y, F43L, G85E, I60T, R81P, and R220Q (Table 1). All of them cause the potential to become more negative which might have an effect on the interactions with other proteins (Fig. 3b). Surface charge-charge relationships are also important in maintaining the stability of the protein [Strickler et al., 2006].

Functional Mutations

V219 is located at the primary specificity site (S1) of HNE (Fig. 3a), and its substitution into I is likely to have an effect on the specificity of the enzyme. R220 is one of the amino acids that forms the specificity pocket (Fig. 3a). It is mutated into Q in cyclic and congenital neutropenia. The active site of HNE is surrounded by a horseshoe-like surface arrangement of 19 arginines, which contributes to the preference of the enzyme to bind linear sulfated polysaccharides [Bode et al., 1989]. A substitution of one of these arginines by a polar, noncharged residue may have an effect on the specificity as well. Residues H53 and C55 are located next to F54, an essential ligand binding site residue. Mutations H53L and C55Y may alter the conformation of F54, thereby leading into specificity change.

C71, which lies next to H70, a residue involved in forming the specificity subsite (S2), contributes to the stability of the protein by forming a disulfide bond with C55. The substitution of C71 with R or S leads into the breakage of the disulfide bond, which is likely to have an effect on at least local structure and the orientation of H70, thereby altering the substrate specificity and affinity of the enzyme.

The C-terminal and N-terminal stretches of HNE (amino acids 1–27 in the N-terminus and 248–267 in the C-terminus) are eliminated by proteolysis during the processing of the mature enzyme. M1V is located in the N-terminal signal peptide. Initiation codon mutation is likely to lead into protein expression abnormalities [Bellanné-Chantelot et al., 2004]. There are two missense mutations in the carboxy-terminal domain (P257L, P262L). The functional consequences of these mutations cannot be predicted, but Benson et al. [2003] suggested that, by analogy with canine cyclic neutropenia, the C-terminal part of HNE might interact with an adaptor protein complex 3 (AP3) involved in protein trafficking. The mutations in the C-terminal part might, thus, impair the intracellular trafficking of the protein. The

TABLE 2. Web Resources for Methods

Service	URL
Conseq	http://conseq.bioinfo.tau.ac.il/
ConSurf	http://consurf.tau.ac.il/
CSU	http://bip.weizmann.ac.il/oca-bin/lpccsu
DisEMBL	http://dis.embl.de/
Disopred	http://bioinf.cs.ucl.ac.uk/disopred/disopred.html
Dmutant	http://phyyz4.med.buffalo.edu/hzhou/mutation
DRIPPRED	www.sbc.su.se/~maccallr/disorder/
ELA2base	http://bioinf.uta.fi/ELA2base
FoldX	http://fold-x.embl-heidelberg.de
GlobPlot	http://globplot.embl.de
KiNG	http://kinemage.biochem.duke.edu/software/king.php
IUPred	http://iupred.enzim.hu/
MultiDisp	http://bioinf.uta.fi/cgi-bin/MultiDisp.cgi
Pfam	www.sanger.ac.uk/Software/Pfam/
PMut	http://mmb.pcb.ub.es:8080/PMut/
PolyPhen	http://coot.embl.de/PolyPhen/
PoPMuSiC	http://babylone.ulb.ac.be/popmusic
PROBE	http://kinemage.biochem.duke.edu/software/probe.php
ProCon	http://dna.uta.fi/ProCon/
PyMOL	http://pymol.sourceforge.net/
RankVia Contact	http://bioinf.uta.fi/RankViaContact.html
RONN	www.strubi.ox.ac.uk/RONN
SCide	www.enzim.hu/scide/ide2.html
SCpred	www.enzim.hu/scpred/pred.html
SIFT	http://blocks.fhcrc.org/sift/SIFT.html
SRide	http://sride.enzim.hu/
TANGO	http://tango.embl.de/

C-terminal stretch is not essential for folding, activation, proteolytic activation or granular targeting of HNE [Gullberg et al., 1995].

In order to interact with AP3, HNE may form transient transmembrane conformations [Benson et al., 2003]. The transmembrane domains are predicted to be placed just before the C-terminal prodomain and near the N-terminus, but distinct from the signal peptide [Benson et al., 2003]. The mutations P205R, L206F, G210V, G214R, V219I, and R220Q are located at the putative C-terminal transmembrane domain, and C55Y, A57T, I60T, and A61V are located at the predicted N-terminal transmembrane domain.

Mutations Predicted to Be Deleterious

We have studied a set of mutations that produce a disease phenotype, thus they are all pathogenic. Pmut predicted 22 of the mutations to be neutral, and seven mutations were predicted to be tolerated by SIFT. PolyPhen predicted four of the mutations to be benign, and all these mutations were also predicted to be neutral by PMut. All the pathogenic mutations predicted by Pmut were predicted to have damaging effects by the other methods, except for S46F, which was tolerated by SIFT.

DISCUSSION

Study of the molecular basis of diseases by experimental methods is laborious and time-consuming, and at the structural level often nearly impossible, especially in cases where there are several missense mutations causing the disease. By contrast, precise and useful information about the effects of mutations on protein structure and function can be readily obtained by theoretical methods.

We have previously applied bioinformatics and structural analysis methods to reveal the basis of, e.g., mutations in different Bruton tyrosine kinase (BTK) domains in X-linked agammaglobulinemia (e.g. [Vihinen et al., 1994a,b, 1995, 1997, 1999]), SH2 domain protein 1A (SH2D1A) mutations in X-linked lymphoproliferative disease [Lappalainen et al., 2000], Bloom syndrome protein (BLM) mutations in Bloom syndrome [Rong et al., 2000], mutations in the Wiscott-Aldrich syndrome protein (WAS) protein in Wiskott-Aldrich syndrome [Rong and Vihinen, 2000], mutations in the methyltransferase (DNMT3B) in immunodeficiency, centromeric instability, and facial abnormalities (ICF) syndrome [Lappalainen and Vihinen, 2002], and CD40L mutations in X-linked hyperimmunoglobulin M (hyper-IgM) syndrome (Thusberg, J., and Vihinen, M., unpublished results). We utilize experience gained in all these studies to discuss the effects of mutations and also the use and applicability of different methods.

The effects of 15 ELA2 missense mutations causing cyclic and congenital neutropenia have been studied previously [Bellanné-Chantelot et al., 2004], by looking only at sequence conservation, effects of mutations on protein stability, secondary structure, and ligand interactions. For sequence alignments they used only four homologous sequences, and only one level of conservation (identity) was studied. Structural aspects were studied by analysis of the PDB structures and modeled mutant structures, and two prediction methods (FoldX and PoPMuSiC) were employed for the analysis of the effects of mutations. We investigated all the known cases reported in literature and used more than 30 methods and tools. Our results differ from the results of Bellanné-Chantelot et al. [2004] in that our results suggest the residue 43 be conserved, and residues 46, 53, 84, 127, and 152 not be conserved. At position 43, hydrophobicity is a conserved property, calculated by ProCon. The accuracy of the sequence conservation analysis is much higher when a large set of homologous sequences is used. We also used FoldX and PoPMuSiC, among other methods, for the prediction of the effects of mutations on structural stability of the protein. As a result of using six independent methods for stability prediction, our results differ from the previous results. According to our results from FoldX and PoPMuSiC, they considered a mutation destabilizing when one of the methods predicted so. In contrast, we require at least three of our methods to agree on the effects. As a result, they predicted the mutations P42L, A127P, and P139L to be destabilizing, where we disagree. Similarly, our results indicate that mutations S46F and V82M cause loss of stability, whereas according to Bellanné-Chantelot et al. [2004], they do not.

We and others have applied numerous methods in this kind of studies; however, systematic discussion of the reliability and performance has been missing. We debate the available methods and highlight our experience. As a study case we analyze ELA2 mutations but base our evaluation of the methods also to the earlier studies on numerous diseases. However, this is not a systematic statistical analysis of the reliability of analysis tools. For that purpose hundreds or thousands of mutations would be needed.

Mutation databases, such as our databases for immunodeficiency-causing mutations (www.bioinf.uta.fi/base_root), serve as the basis for bioinformatics research on the effects of mutations and the structural basis of diseases. Conserved amino acids tend to be essential for structure and function, which is why disease-causing mutations often occur at these positions [Miller and Kumar, 2001; Mooney and Klein, 2002]. The probability that a random mutation can cause a genetic disease has been shown to increase with an increase in the degree of site conservation

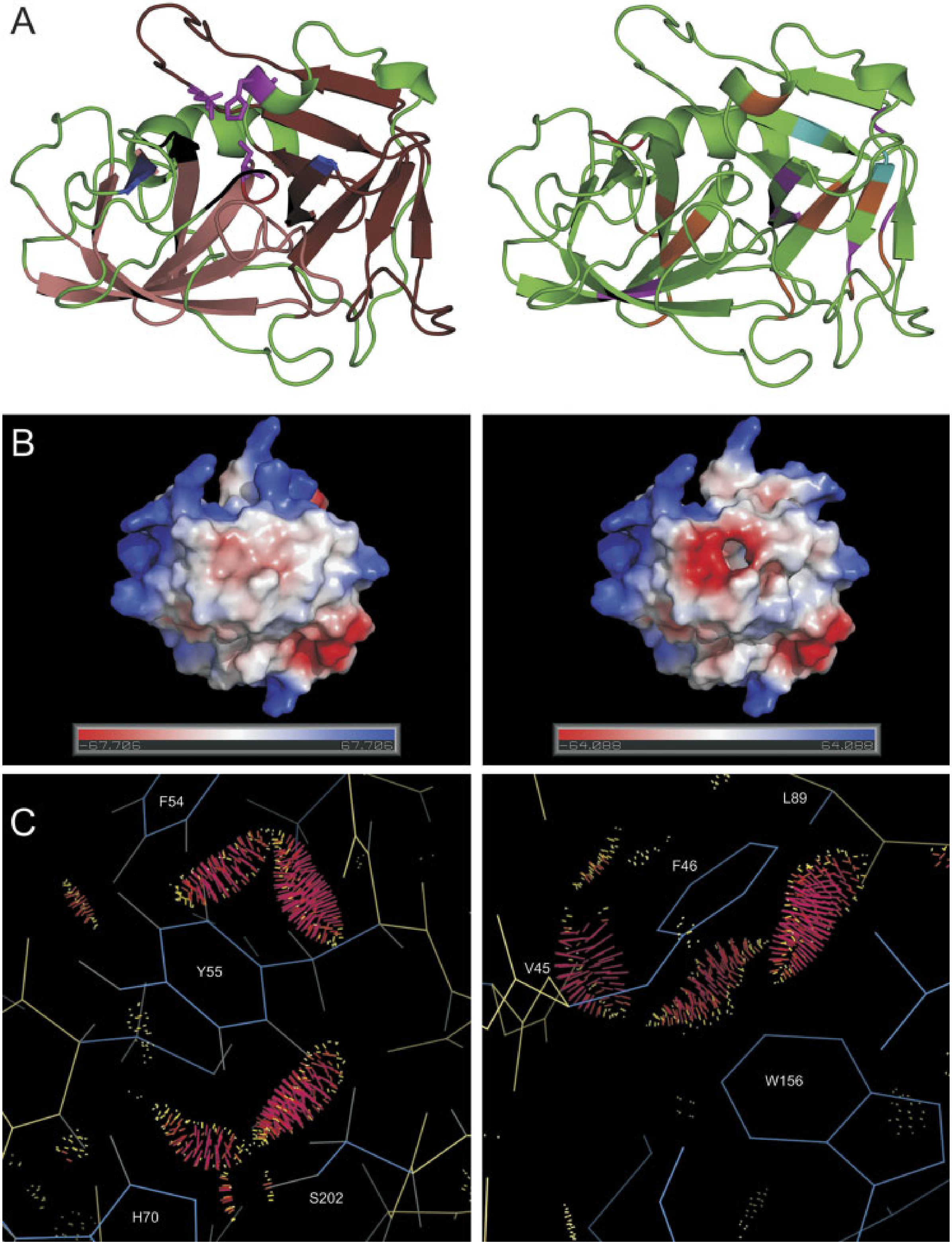


FIGURE 3. **A:** Stereo view of HNE trypsin homology domain. On the left, the two β barrels forming the main fold of the protein are colored pink and brown. Magenta: the catalytic triad. Red: the oxyanion hole. Black: the primary specificity site. Blue: other residues involved in ligand binding. On the right, the HNE missense mutations are colored according to the primary effects of the individual mutations on protein structure and function. Orange: conformational mutations. Magenta: Mutations affecting the stability-maintaining contacts. Cyan: Mutations increasing structural disorder. Red: functional mutations. **B:** Electrostatic surface potentials in wild-type (left) and mutated (right) HNE. In the mutated structures, all 29 missense mutations located in the trypsin homology domain are included. The mutations alter the surface potential from positive to negative. **C:** Substitutions C55Y and S46F cause serious clashes with the neighboring residues.

[Vitkup et al., 2003]. The nature of the amino acid substitution at an invariant site can reveal the effect of the mutation on protein structure or function. The variable positions give insight into the types of amino acids that can be freely exchanged without negatively impacting protein function [Miller and Kumar, 2001]. In the case of HNE, most missense mutations occur at the conserved positions (Table 1; Fig. 1). Sequence analysis can reveal many details about the mutated sites, sequence conservation and allowed/disallowed residues at the alignment positions. There are several methods available for studying and visualizing the degree of sequence conservation. Ready-made sequence alignments for protein families and domains can be obtained from the Pfam database [Bateman et al., 2004]. The MultiDisp program (Riikonen, P., Horvath, T., and Vihinen, M., unpublished results) visualizes sequence alignments for the detection of conserved amino acids and highlights the chemical nature of the residues (Fig. 1). MultiDisp contains numerous options for visualizing sequence similarity and differences, physicochemical properties of the alignment positions and position-wise conservation analysis. Conseq and Consurf servers visualize sequence alignments by a color scheme indicating the degree of conservation in the aligned positions. Consurf also visualizes the conserved positions in the query protein structure. Conseq features a prediction of the buried and exposed positions on globular proteins based on a neural network prediction scheme. This can be a useful feature if structural information is not available. The program al2co calculates conservation indices for sequence alignments by three conceptually different approaches (entropy-based, variance-based and matrix score-based) and by both weighted and unweighted amino acid frequencies defined by the user [Pei and Grishin, 2001]. The determination of conservation within alignment positions is not straightforward, although often somewhat arbitrary frequency percentage has been used as a threshold. The situation is more complex and therefore only statistical methods can reveal weaker conservation [Ahola et al., 2003, 2004]. Since we studied effects of mutations, protein family information was used to reveal sites with a high level of conservation as well as to investigate amino acid substitutions that appeared in some family members.

ProCon identifies amino acid conservation at three different levels (identity, conservation of the chemical and physical nature of amino acids, and covariant conservation). The conserved positions can be visualized both at the sequence level and in the protein structure. The method is based on entropy calculations for amino acid positions in multiple sequence alignments [Shen and Vihinen, 2004]. All these programs are very useful for protein family analysis and sequence conservation study at position level. The results also indicate which residues or properties of residues are crucial for certain sequence positions. These methods are even more useful when a structure of at least one family member is known. A single analysis cannot reveal all sequence features, therefore several tools should be applied or the different features of, e.g., MultiDisp should be used.

Proteins fold to their typical three-dimensional structures, but the structure is dynamic and in constant motion at different amplitudes and frequencies. Many protein structures contain segments which do not have a well-organized structure and some proteins have even global disorder, i.e., do not fold in an ordered way. Sequence-based prediction methods for protein structural disorder and aggregation propensities are based on protein amino acid composition and amino acid physicochemical properties and energy profiles, missing X-ray coordinates and specific sequence patterns. The DISOPRED server [Ward et al., 2004a] uses the DISOPRED2 dynamic disorder prediction method [Ward et al., 2004b] for amino acid sequences. The method identifies disorder defined as missing coordinates in high-resolution X-ray crystal structure electron density maps, by which the method has been trained. The method runs PSI BLAST searches [Altschul et al., 1997] and estimates the probability of a residue being disordered by encoding each residue by the profile for a window of 15 positions in the sequence and classifying the residues using a neural network [Ward et al., 2004b]. The major drawback of the DISOPRED method is that the missing coordinates, which have been used to define the concept of disorder, can also arise as an artifact of the crystallization process.

DisEMBL is based on a neural network trained for predicting several definitions of disorder. Like the DISOPRED method, DisEMBL predicts disorder in protein sequences using missing coordinates in X-ray structures. As additional criteria in the DisEMBL method, the disordered regions must reside within loops or coils. B factors are also taken into account, so that highly dynamic loops are considered to be disordered [Linding et al., 2003a].

Globplot, a tool for recognizing globular and disordered regions within amino acid sequences, is based on Russell/Linding secondary structure-forming propensities [Linding et al., 2003b], where propensities presented by Deleage and Roux [1987] are combined with propensities of amino acids to be in regular secondary structures as defined by dictionary of protein secondary structure (DSSP) [Kabsch and Sander, 1983] or outside of them [Linding et al., 2003b]. In the DRIP-PRED method, self-organizing maps (SOM) 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 sequence regions are predicted to be disordered (www.sbc.su.se/~maccallr/disorder). The problem with this approach is that PDB is biased and does not contain all types of structures. It is likely that the disorder tools overpredict and overestimate the effect of disorder at least in globular proteins.

IUPred 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 disordered [Dosztányi et al., 2005a,b]. A regional order neural network, RONN, predicts disorder by comparing the input sequence to other sequences of known folding state: ordered, disordered, or a mixture of both [Yang et al., 2005].

In the program TANGO, aggregation propensities for proteins are estimated by amino acid physicochemical properties and competition between different structural conformations: α-helix, β-turn, β-sheet aggregates, and the folded state [Fernandez-Escamilla et al., 2004; Linding et al., 2004]. The method of Chiti et al. [2003] predicts changes in aggregation rates induced by mutations using the change in hydrophobicity, in the propensity to convert from α-helical to β-sheet structure, and in overall charge as parameters.

Missense mutations may increase the structural disorder and aggregation propensity in proteins, so that the protein fails to fold correctly and there are regions lacking regular secondary structure. Disorder may also be secondary and arise as a result of structural alterations; e.g., due to sterical clashes of amino acids. Such structural changes into the protein may have detrimental effects on protein function. Protein disorder has been demonstrated to play a central role in diseases mediated by protein misfolding and aggregation, for example Alzheimer's and Huntington's diseases and amyloidosis [Schweers et al., 1994; Kaplan et al., 2003; Bates, 2003; Grateau et al., 2005]. Of the missense mutations in HNE, 25% are predicted to cause increase in structural disorder by most of the methods (Table 1). Examples of the results provided by the disorder and aggregation propensity prediction methods are provided in Fig. 4. Although there are several methods available for disorder prediction, they seldom agree on the effects of mutations, which is to be expected because the concepts of the methods are so different and so far the size of the training sets is limited.

Compromised folding and decreased stability of the protein are the major molecular pathogenic consequences of missense mutations [Bross et al., 1999]. Structural stability is maintained by interactions between amino acids, and the hydrophobic core of the protein has a central role in protein structural integrity. RankViaContact is a program that calculates contact energies for amino acids in protein structures and also visualizes the structure with the residues with highest or lowest contact energies highlighted [Shen and Vihinen, 2004]. Mutations that affect amino acids forming strong contacts and disulfide bonds or ionic interactions can be considered destabilizing. Amino acids located in the core of the protein, with a negligible or very small solvent accessible surface area, typically form several hydrophobic interactions essential for the fold and the stability of the protein structure [Shortle et al., 1990; Serrano et al., 1992; Matthews, 1995; Sandberg et al., 1995]. Mutations in such amino acids may cause detrimental changes to the structure-maintaining contacts. Six missense mutations in HNE affect the residues forming the strongest structure maintaining contacts (Fig. 4c). The more buried a residue, the more likely are mutations of the site to be pathogenic [Vitkup et al., 2003]. In addition, the introduction of charged side chains into the hydrophobic core is known to destabilize protein structure [Chasman and Adams, 2001].

The methods predicting protein stability are based on calculating the free energies of folding (FoldX), mutation-induced thermodynamic stability changes and changes in free energy of folding (PoPMuSic). The program DMUTANT utilizes distance-dependent, residue-specific all-atom potentials [Zhou and Zhou, 2002]. The method uses finite ideal-gas reference state. It has been tested with 895 disease-causing mutations in proteins for which the three dimensional structure has been determined. The correlation coefficient for experimentally studied and predicted changes was shown to be 0.67 [Zhou and Zhou, 2002]. Stabilizing amino acids can be predicted based on long-range interactions in protein structures (Sride, Scide), and hydrophobicity and conservation of amino acid residues (Sride) [Dosztányi et al., 2003a; Magyar et al., 2005]. The Scpred method is based on differences in sequential neighborhood [Dosztányi et al., 2003b]. Of the 32 analyzed HNE mutations, 18 seem to decrease protein stability (Table 1). SIFT and Pmut evaluate the pathogenicity of mutations by the occurrence of amino acids in sequence alignments (SIFT and Pmut), and amino acid properties, secondary structure and accessibility predictions (Pmut) [Ng and Henikoff, 2001; Ferrer-Costa et al., 2005]. PolyPhen, another method for distinguishing damaging mutations from neutral ones, is based on knowledge of the functional sites of the protein, positional residue variation in sequence alignments, and the 3D structure of the protein [Sunyaev et al., 2001; Ramensky et al., 2002]. When studying the effects of disease-causing mutations, these predictors can be useful in determining the reason for the pathogenicity of the mutation, even though the actual predictions (pathogenic/tolerated) of these programs mostly do not agree with each other or with results from other analyses. The reliability of SIFT has been shown to be 78% [Saunders and Baker, 2002].

The effects of mutations on the protein structure, especially the fit of the introduced residues, can be studied quantitatively by the Autobondrot function of the program PROBE, which rotates each side chain χ angle on the selected residue and scores each combination of amino acid χ angles by analyzing side chain packing interactions [Word et al., 1999b, 2000]. The mutated structures in which the introduced amino acid side-chain χ angles are optimized according to the best rotamer ranked by the PROBE score are best visualized with KiNG. Visualization of the structure allows qualitative analysis of the contacts between amino acids. This is still a good approach for evaluating the effects of mutations, but requires an expert in protein structures. Of the HNE missense mutations, 90% cause major structural changes to the protein, according to the Probe score for the best rotamer and analysis of the side chain clashes caused by the mutations (Table 1; Fig. 3a and c). Side chains with a low score do not fit into the structure in any conformation, causing changes to the structure already during the folding process.

Qualitative electrostatic surface potentials can be calculated, e.g., with the program PyMOL [DeLano, 2002] or Delphi [Klapper et al., 1986; Gilson et al., 1988; Gilson and Honig, 1988]. In addition to the shape, also charge properties are important, e.g., for ligand (substrate, cofactor, inhibitor, etc.) interactions. Mutations in many cases significantly reduce or even reverse the local electrostatics, with consequent effects on binding, specificity, etc. Changes in the electrostatic potential affect the properties of a protein in many ways. Electrostatics is a significant factor in protein folding and stability, and it has a crucial effect on protein interactions. Changes in the electrostatic surface potentials appear in 8 of the 29 analyzed mutations (Table 1; Fig. 3b).

Missense mutations causing cyclic and congenital neutropenia have rather a structural than functional character, many of them affecting the stability and structural assembly of HNE. Of the 29 missense mutations located in the trypsin homology domain, 20 cannot fit into the structure in any conformation or lead to clashes with neighboring residues. A total of 19 mutations are predicted to cause decrease of structural stability, and six of the mutations are likely to reduce the essential stability maintaining contacts between residues. Disorder and aggregation propensities are increased by 12 mutations (Table 1). Only one mutation directly affects a specific functional site of the enzyme, and a few mutations are located in the vicinity of the functional sites (Fig. 3a). Some of the mutations have a putative role in intracellular protein trafficking (Table 1).

Genotype–phenotype correlations are particularly interesting in the case of HNE, because two different disease phenotypes arise as an effect of the mutations. Cyclic and congenital neutropenia are caused by mutations in the same protein, sometimes even an individual substitution can cause both forms of the disease in different families (Table 1). In this study we were not able to a find clear correlation in the disease phenotype and the effects of missense mutations on HNE structure and function.

We have utilized 31 methods for predicting the effects of missense mutations at the sequence and structural level. Sequence level predictions include methods for analyzing the degree of conservation of the positions in multiple sequence alignments, predicting the effects of mutations on protein disorder and aggregation propensities, evaluating the pathogenicity of the missense mutations, and predicting the effects of mutations on the stability of the protein. On the structural level, we implemented methods for optimizing the torsional angles for the introduced side chains, analyzing the interresidue contacts

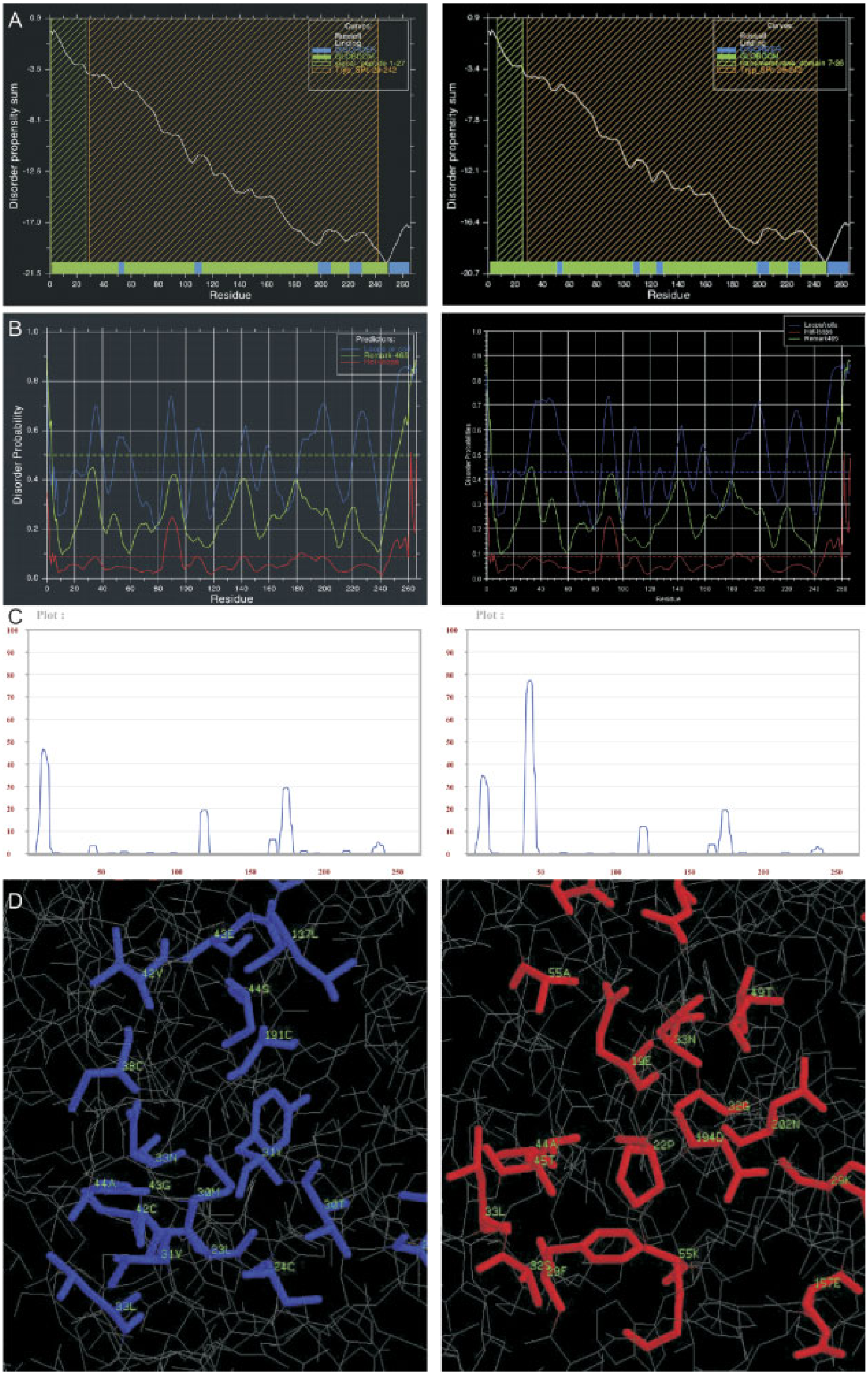


FIGURE 4. **A:** Examples of analysis of missense mutation effects. The mutation A127P is predicted to increase disorder in the protein by Globplot. **B:** The mutation L47P is predicted to increase disorder in the protein by Disopred. **C:** P42L increases HNE aggregation propensity predicted by TANGO. **D:** RankViaContact visualizes important contact-forming residues. Left: residues involved in forming the structure-maintaining contacts in the core of the protein. Right: residues forming strong contacts on the surface of the protein. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

essential for protein structural stability, visualization of the structure and predicting the effects of mutations on the protein electrostatic surface potential. There are still more bioinformatics methods which can be used for this kind of analyses. Homology modeling can be used to predict the three-dimensional protein structure when no experimental structure is available. Then many of the structure-based analyses can be performed by using the model. Molecular dynamics simulations can be used to study numerous dynamic structural features, catalytic activity and, e.g., binding affinity of wild-type and mutated proteins.

Effects of missense mutations on protein structure and function can be analyzed even when the structure for the protein has not been determined. However, structure-based analysis methods provide more information on the effects and are likely more reliable. Sequence-level prediction methods are limited to analysis of residue conservation in protein families, and prediction of aggregation and disorder propensities. Some of the methods for predicting the stability of a protein can also be used with sequence data only. The results obtained by sequence-based methods can be evaluated based on the structural data, and by combining the results from multiple analysis methods makes the predictions more reliable. The effects of mutations on the residue-residue contacts maintaining the stability of the protein and the ability of the introduced side chains to fit into the structure are possible to analyze only by structure-based methods. These aspects are crucial for analyzing protein structure–function relationships and effects of mutations, because the majority of disease-causing mutations have structural, rather than functional effects [Wang and Moult, 2001; Mooney and Klein, 2002]. 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 starting points. For example, structural changes may originate from changes in side chain size, charge, hydropathy, altered contact-forming properties, aggregation, or introduced disorder.

One must be familiar with the theory and limitations of the various prediction methods in order to be able to interpret the results correctly. In addition, deep knowledge of protein structures and chemistry are essential. This and our previous studies have indicated that it is beneficial to use several analysis methods whenever possible. There are numerous bioinformatics methods available for a number of tasks. Even when the three dimensional structure is solved at high resolution it is beneficial to use some of the sequence-based methods, especially for conservation analysis within the protein family. For some other features the structure provides a superior starting point. All these methods provide putative explanations for the diseases. Our experience indicates that careful choice and understanding of the methods and their limitations is important to avoid overprediction and to provide insight to the causes of diseases.

ACKNOWLEDGMENT

We thank Kathryn Rannikko for language revision.

REFERENCES

Adkinson AM, Raptis SZ, Kelley DG, Pham CT. 2002. Dipeptidyl peptidase I activates neutrophil-derived serine proteases and regulates the development of acute experimental arthritis. J Clin Invest 109:363–371.

Aghamohammadi A, Parvaneh N, Kanegana H, Moin M, Amirzargar AA, Farhoudi A, Pourpak Z, Movahedi M, Gharagozlou M, Rezaei N, Futatani T, Miyawaki T. 2004. Screening of the Bruton tyrosine kinase (BTK) gene mutations in 13 Iranian patients with presumed X-linked agammaglobulinemia. Iran J Allergy Asthma Immunol 3:175–179.

Ahola V, Uusipaikka E, Aittokallio T, Vihinen M. 2003. Efficient estimation of emission probabilities in profile hidden Markov models. Bioinformatics 19:2359–2368.

Ahola V, Aittokallio T, Vihinen M, Uusipaikka E. 2004. Statistical multiple comparison methods for identifying conserved residues in protein sequence alignment. Stat Appl Genet Mol Biol 3:28.

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402.

Bach-Gansmo ET, Halvorsen S, Godal HC, Skjønsberg OH. 1996. D-dimers are degraded by human neutrophil elastase. Thromb Res 82: 177–186.

Barrett AJ, Rawlings ND. 1995. Families and clans of serine proteases. Arch Biochem Biophys 318:247–250.

Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, Khanna A, Marshall M, Moxon S, Sonnhammer EL, Studholme DJ, Yeats C, Eddy SR. 2004. The Pfam protein families database. Nucleic Acids Res 32:D138–D141.

Bates G. 2003. Huntingtin aggregation and toxicity in Huntington's disease. Lancet 361:1642–1644.

Belaaouaj A, Kim KS, Shapiro SD. 2000. Degradation of outer membrane protein A in *Escherichia coli* killing by neureophil elastase. Science 289:1185–1187.

Bellanné-Chantelot C, Clauin S, Leblanc T, Cassinat B, Rodrigues-Lima F, Beaufils S, Vaury C, Barkaoui M, Fenneteau O, Maier-Redelsperger M, Chomienne C, Donadieu J. 2004. Mutations in the ELA2 gene correlate with more severe expression of neutropenia: a study of 81 patients from the French Neutropenia Register. Blood 11:4119–4125.

Benson KF, Li F-Q, Person RE, Albani D, Duan Z, Wechsler J, Meade-White K, Williams K, Acland GM, Niemeyer G, Lothrop CD, Horwitz M. 2003. Mutations associated with neutropenia in dogs and humans disrupt intracellular transport of neutrophil elastase. Nat Genet 35: 90–96.

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.

Bieth JG. 1998. Leukocyte elastase. In: Barrett AJ, Rawlings ND, Woessner JF, editors. Handbook of proteolytic enzymes. San Diego: Academic Press. p 54–60.

Bode W, Meyer E Jr, Powers JC. 1989. Human leukocyte and porcine pancreatic elastase: X-ray crystal structures, mechanism, substrate specificity, and mechanism-based inhibitors. Biochemistry 28: 1951–1963.

Bode W, Wei AZ, Huber R, Meyer E, Travis J, Neumann S. 1986. X-ray structure of the complex of human leukocyte elastase (PMN elastase) and the third domain of the turkey ovomucoid inhibitor. EMBO J 5:2453–2458.

Boxer LA, Morganroth ML. 1987. Neutrophil function disorders. Dis Mon 33:681–780.

Brinkmann V, Reichard U, Goosmann C, Fauler B, Uhlemann Y, Weiss DS, Weinrauch Y, Zychlinsky A. 2004. Neutrophil extracellular traps kill bacteria. Science 303:1532–1535.

Bross P, Corydon TJ, Andersen BS, Jørgensen MM, Bolund L, Gregersen N. 1999. Protein misfolding and degradation in genetic diseases. Hum Mutat 14:186–198.

Chasman D, Adams RM. 2001. Predicting the functional consequences of nonsynonymous single nucleotide polymorphisms: structure based assessment of amino acid variation. J Mol Biol 307:683–706.

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.

Dale DC, Hammond WPT. 1988. Cyclic neutropenia: a clinical review. Blood Rev 2:178–185.

Dale DC, Person RE, Bolyard AA, Aprikyan AG, Bos C, Bonilla MA, Boxer LA, Kannourakis G, Zeidler C, Welte K, Benson KF, Horwitz M. 2000. Mutations in the gene encoding neutrophil elastase in congenital and cyclic neutropenia. Blood 96:2317–2322.

DeLano WL. 2002. The PyMOL molecular graphics system. San Carlos, CA: DeLano Scientific. Available at: http://www.pymol.org. Last date accessed: August 1, 2006.

Deleage G, Roux B. 1987. An algorithm for protein secondary structure prediction based on class prediction. Protein Eng 1:289–294.

Dewald B, Rindler-Ludwig R, Bretz U, Baggiolini M. 1975. Subcellular localization and heterogeneity of neutral proteases in neutrophilic polymorphonuclear leukocytes. J Exp Med 141:709–723.

Dobson CM. 2003. Protein folding and misfolding. Nature 426:884–890.

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, Cserzá M, Fiser A, Simon I. 2003b. Servers for sequence-structure relationship analysis and prediction. Nucleic Acids Res 31:3359–3363.

Dosztányi Z, Csizmók V, Tompa P, Simon I. 2005a. IUPred: web server for the prediction of intrinsically unstructured proteins. Bioinformatics 21: 3433–3434.

Dosztányi Z, Csizmók 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.

Fernandez-Escamilla AM, Rousseau F, Schymkowitz J, Serrano L. 2004. Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins. Nature Biotechnol 22:1240–1241.

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.

Forloni G, Terreni L, Bertani I, Fogliarino S, Invernizzi R, Assini A, Ribizzi G, Negro A, Calabrese E, Volonte MA, Mariani C, Franceschi M, Tabaton M, Bertoli A. 2002. Protein misfolding in Alzheimer's and Parkinson's disease: genetics and molecular mechanisms. Neurobiol Aging 23:957–976.

Gilis D, Rooman M. 2000. PoPMuSiC, an algorithm for predicting protein mutant stability changes: application to prion proteins. Protein Eng 13: 849–856.

Gillis S, Furie BC, Furie B. 1997. Interactions of neutrophils and coagulation proteins. Semin Hematol 34:336–342.

Gilson M, Honig B. 1988. Calculation of the total electrostatic energy of a macromolecular system: solvation energies, binding energies, and conformational analysis. Proteins Struct Funct Genet 4:7–18.

Gilson M, Sharp K, Honig B. 1988. Calculating electrostatic interactions in biomolecules: methods and error assessment. J Comput Chem 9: 327–335.

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.

Grateau G, Verine J, Delpech M, Ries M. 2005. Amyloidosis: a model of misfolded protein disorder. Med Sci 21:627–633.

Gromiha MM, Pujadas G, Magyar C, Selvaraj S, Simon I. 2004. Locating the stabilizing residues in α/β barrel proteins based on hydrophobity, long-range interactions, and sequence conservation. Proteins 55: 316–329.

Gullberg U, Lindmark A, Lindgren G, Persson AM, Nilsson E, Olsson I. 1995. Carboxyl-terminal prodomain-deleted human leukocyte elastase and cathepsin G are targeted to granules and enzymatically activated in the rat basophilic/mast cell line RBL. J Biol Chem 270:12912–12918.

Haurie C, Dale DC, Mackey MC. 1998. Cyclical neutropenia and other periodic haematological disorders: a review of mechanisms and mathematical models. Blood 92:2629–2640.

Horwitz M, Benson KF, Person RE, Aprikyan AG, Dale DC. 1999. Mutations in ELA2, encoding neutrophil elastase, define a 21-day biological clock in cyclic hematopoiesis. Nat Genet 23:433–436.

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

Kaplan B, Ratner V, Haas E. 2003. α-Synuclein: Its biological function and role in neurodegenerative diseases. J Mol Neurosci 20:83–92.

Klapper I, Hagstrom R, Fine R, Sharp K, Honig B. 1986. Focusing of electric fields in the active site of Cu-Zn superoxide dismutase: effects of ionic strength and amino acid modification. Proteins Struct Funct Genet 1:47–59.

Kraut J. 1977. Serine proteases: structure and mechanism of catalysis. Ann Rev Biochem 46:331–358.

Kwasigroch JM, Gilis D, Dehouck Y, Rooman M. 2002. PoPMuSiC, rationally designing point mutations in protein structures. Bioinformatics 18:1701–1702.

Lacroix E, Viguera AR, Serrano L. 1998. Elucidating the folding problem of α-helices: Local motifs, long-range electrostatics, ionic-strength dependence and prediction of NMR parameters. J Mol Biol 284:173–191.

Lappalainen I, Giliani S, Franceschini R, Bonnefoy J-Y, Duckett C, Notarangelo LD, Vihinen M. 2000. Structural basis for SH2D1A mutations in X-linked lymphoproliferative disease (XLP). Biochem Biophys Res Commun 269:124–130.

Lappalainen I, Vihinen M. 2002. Structural basis of ICF-causing mutations in the methyltransferase domain of DNMT3B. Prot Eng 15: 1005–1014.

Lehrer RI, Ganz T, Selsted ME, Babior BM, Curnutte JT. 1988. Neutrophils and host defense. Ann Intern Med 109:127–142.

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.

Linding R, Schymkovitz J, Rousseau F, Diella F, Serrano L. 2004. A comparative study of the relationship between protein structure and β-aggregation in globular and intrinsically disordered proteins. J Mol Biol 342:345–353.

Lovell SC, Word JM, Richardson JS, Richardson DC. 2000. The penultimate rotamer library. Proteins 40:389–408.

Lovell SC, Davis IW, Arendall WB, de Bakker PIW, Word JM, Prisant MG, Richardson JS, Richardson DC. 2003. Structure validation by Cα geometry: ϕ, ψ and cβ deviation. Proteins 50:437–450.

MacCallum RM. 2006. Order/disorder prediction with self organising maps. Available at: http://www.sbc.su.se/~maccallr/disorder. Last accessed: August 1, 2006.

Magyar C, Gromiha MM, Pujadas G, Tusnady GE, Simon I. 2005. Sride: a server for identifying stabilizing residues in proteins. Nucleic Acids Res 33:W303–W305.

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.

Mooney SD, Klein TE. 2002. The functional importance of disease-associated mutation. BMC Bioinformatics 3:24.

Muños V, Serrano L. 1994. Elucidating the folding problem of helical peptides using empirical parameters. Nat Struct Biol 1:399–409.

Neurath H. 1986. The versatility of proteolytic enzymes. J Cell Biochem 32:35–49.

Ng PC, Henikoff S. 2001. Predicting deletorious amino acid substitutions. Genome Res 11:863–874.

Palmer SE, Stephens K, Dale DC. 1996. Genetics, phenotype and natural history of autosomal dominant cyclic hematopoiesis. Am J Med Genet 66:413–422.

Pei J, Grishin NV. 2001. AL2CO: calculation of positional conservation in a protein sequence alignment. Bioinformatics 17:700–712.

Piirilä H, Vihinen M. 2006. Immunodeficiency mutation databases (IDbases). Hum Mutat 27:1200–1208.

Poussu E, Vihinen M, Paulin L, Savilahti H. 2004. Probing the α-complementing domain of *E.* coli β-galactosidase with use of an insertional pentapeptide mutagenesis strategy based on Mu in vitro DNA transposition. Proteins 54:681–692.

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

Rong S-B, Väliaho J, Vihinen M. 2000. Structural basis of Bloom syndrome (BS) causing mutations in the BLM helicase domain. Mol Med 6: 155–164.

Rong S-B, Vihinen M. 2000. Structural basis of Wiscott-Aldrich syndrome (WAS) causing mutations in the WH1 domain. J Mol Med 78: 530–537.

Salvesen G, Enghild JJ. 1990. An unusual specificity in the activation of neutrophil serine proteinase zymogens. Biochemistry 29:5304–5308.

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.

Saunders CT, Baker D. 2002. Evaluation of structural and evolutionary contributions to deleterious mutation prediction. J Mol Biol 322: 891–901.

Sanders CR, Myers JK. 2004. Disease-related misassembly of membrane proteins. Annu Rev Biophys Biomol Struct 33:25–51.

Schweers O, Schönbrunn-Hanebeck E, Marx A, Mandelkow E. 1994. Structural studies of tau protein and Alzheimer paired helical filaments show no evidence for β-structure. J Biol Chem 269:24290–24297.

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.

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.

Shortle D, Stites WE, Meeker AK. 1990. Contributions of the large hydrophobic amino acids to the stability of staphylococcal nuclease. Biochemistry 29:8033–8041.

Serrano L, Kellis JT Jr, Cann P, Matouschek 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.

Sobolev V, Sorokine A, Prilusky J, Abola EE, Edelman M. 1999. Automated analysis of interatomic contacts in proteins. Bioinformatics 15:327–332.

Steward RE, MacArthur MW, Laskowski RA, Thornton JM. 2003. Molecular basis of inherited diseases: a structural perspective. Trends Genet 19:505–513.

Strickler SS, Gribenko AV, Gribenko AV, Keiffer TR, Tomlinson J, Reihle T, Loladze VV, Makhatadze GI. 2006. Protein stability and surface electrostatics: a charged relationship? Biochemistry 45:2761–2766.

Sunyaev S, Ramensky V, Koch I, Lathe W, Kondrashov A, Bork P. 2001. Prediction of deleterious human alleles. Hum Mol Gen 10:591–597.

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, 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, Zvelebil JJM, Zhu Q, Brooimans RA, Ochs HD, Zegers BJM, 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.

Vogt G, Chapgier A, Yang K, Chuzhanova N, Feinberg J, Fieschi C, Boisson-Dupuis S, Alcais A, Filipe-Santos O, Bustamante J, de Beaucourdey L, Al-Mohsen I, Al-Haijar S, Al-Ghonaium A, Adimi P, Mirsaeidi M, Khalilzadeh S, Rosenzweig S, de la Calle Martin O, Bauer TR, Puck JM, Ochs HD, Furthner D, Engelhorn C, Belohradsky B, Mansouri D, Holland SM, Schreiber RD, Abel L, Cooper DN, Soudais C, Casanova J-L. 2005. Gains of glycosylation comprise an unexpectedly large group of pathogenic mutations. Nat Gen 37:692–700.

Ward JJ, McGuffin LJ, Bryson K, Buxton BF, Jones DT. 2004a. The DISOPRED server for the prediction of protein disorder. Bioinformatics 20:2138–2139.

Ward JJ, Sodhi JS, McGuffin LJ, Buxton BF, Jones DT. 2004b. Prediction and functional analysis of native disorder in proteins from three kingdoms of life. J Mol Biol 337:635–645.

Weiss SJ. 1989. Tissue destruction by neutrophils. N Engl J Med 320: 365–376.

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

Weinrauch Y, Drujan D, Shapiro SD, Weiss J, Zychlinsky Z. 2002. Neutrophil elastase targets virulence factors of enterobacteria. Nature 417:91–94.

Wiedow O, Muhle K, Streit V, Kameyoshi Y. 1996. Human eosinophils lack human leukocyte elastase. Biochim Biophys Acta 1315:185–187.

Wintroub BU, Coblyn JS, Kaempfer CE, Austen KF. 1980. Cleavage of fibrinogen by the human neutrophil neutral peptide-generating protease. Proc Natl Acad Sci USA 77:5448–5452.

Word JM, Lovell SC, La Bean TH, Taylor HC, Zalis ME, Presley BK, Richardson JS, Richardson DC. 1999a. Visualizing and quantifying molecular goodness of fit: small probe contact dots with explicit hydrogen atoms. J Mol Biol 285:1711–1733.

Word JM, Lovell SC, La Bean TH, Taylor HC, Zalis ME, Presley BK, Richardson JS, Richardson DC. 1999b. Visualizing and quantifying molecular goodness of fit: small probe contact dots with explicit hydrogen atoms. J Mol Biol 285:1711–1733.

Word JM, Bateman RC Jr, Presley BK, Lovell SC, Richardson DC. 2000. Exploring steric constraints on protein mutations using MAGE/PROBE. Protein Sci 9:2251–2259.

Yang ZR, Thompson R, McMeil 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.

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.

Protein Engineering, Design & Selection vol. 20 no. 3 pp. 133–141, 2007
Published online February 16, 2007 doi:10.1093/protein/gzm004

The structural basis of hyper IgM deficiency – CD40L mutations

J.Thusberg[1] and M.Vihinen[1,2,3]

[1]Institute of Medical Technology, FI-33014, University of Tampere, Finland,
[2]Research Unit, Tampere University Hospital,
FI-33520 Tampere, Finland

[3]To whom correspondence should be addressed. Institute of Medical Technology, FI-33014, University of Tampere, Finland. Email: mauno.vihinen@uta.fi

X-linked hyper-IgM syndrome (XHIGM) is a primary immunodeficiency characterised by an inability to produce immunoglobulins of the IgG, IgA and IgE isotypes. It is caused by mutations of CD40 ligand (CD40L, CD154), expressed on T-lymphocytes. The interaction of CD40L on T-cells and its receptor CD40 on B-cells is essential for lymphocyte signalling leading to immunoglobulin class switching and B-cell maturation. To understand the structural basis for XHIGM, we utilised bioinformatics methods to analyse all the known CD40L missense mutations at both the sequence and structural level. Our results demonstrate that the 35 different missense mutations have diverse effects on CD40L structure and function, affecting structural disorder and aggregation tendencies, stability maintaining contacts and electrostatic properties. Several mutations also affect residues essential in receptor binding and trimerisation. Experimental study of effects of mutations is laborious and time-consuming and at the structural level often almost impossible. By contrast, precise and useful information about effects of mutations on protein structure and function can readily be obtained by theoretical methods. In this study, all the XHIGM causing missense mutations could be explained in terms of CD40L structure and function. Thus, the molecular basis of the syndrome could be elucidated.

Keywords: bioinformatical analysis/disease-causing mutations/immunodeficiencies/structural basis of disease/structure–function relationships

Introduction

X-linked hyper-IgM syndrome (XHIGM; OMIM 308230) is a rare and severe primary immunodeficiency characterised by the absence or low levels of IgG, IgA and IgE, normal or elevated IgM level in serum, and defective immunoglobulin class switch recombination (Notarangelo *et al.*, 1992; Fuleihan *et al.*, 1993). XHIGM patients are highly susceptible to recurrent bacterial infections and they are prone to autoimmune diseases and neutropenia (Levy *et al.*, 1997). The syndrome is caused by mutations of CD40 ligand (CD40L, CD154), expressed on T-cells, and the inability of the mutated protein to bind to its receptor CD40 on B-cells (Aruffo *et al.*, 1993).

CD40L, a member of the tumour necrosis factor (TNF) family of cytokines, is a 39 kDa Type II membrane glycoprotein expressed primarily on activated CD4$^+$ T-cells (Noelle *et al.*, 1992). The CD40L monomer consists of four distinct structural domains: an N-terminal intracellular tail (amino acids 1–22), a short transmembrane domain (amino acids 23–46), a portion that forms the extracellular unique domain (amino acids 47–122) and the extracellular, C-terminal TNF homology (TNFH) domain (amino acids 123–261). The crystal structure of the CD40L TNFH domain has been determined to 2.0 Å resolution (Karpusas *et al.*, 1995).

TNFH domain superfamily members have a highly conserved jelly roll type structure, consisting of two β sheets that have a Greek key topology. The TNFH domains are responsible for receptor binding. The sequence identity between family members is ∼20–30% (Bodmer *et al.*, 2002).

The CD40L–CD40 interaction is essential in B-cell activation and antibody isotype switching (Kroczek *et al.*, 1994). Isotype switching by B-cells stimulated by T-dependent signals requires both the ligation of CD40 and a second signal provided by a T-cell derived cytokine (Coffman *et al.*, 1993). CD40 is constantly expressed on B-cells (Clark and Ledbetter, 1986), whereas CD40L is expressed only after class II major histocompatibility factor (MHC)–T-cell receptor (TCR) interaction and T-cell activation (Armitage *et al.*, 1992). The expression of CD40L is also regulated in an autologous manner, so that the interaction of the ligand with its receptor upregulates its own expression (Pinchuk *et al.*, 1996).

The ligand–receptor interaction triggers a signalling cascade leading to the activation of several genes involved in B-cell proliferation and antibody production (Allen *et al.*, 1993), and the downregulation of genes whose expression has been shown to lead to cell cycle arrest (Dadgostar *et al.*, 2002). The expression of B7 proteins on the B-cell surface is also stimulated by the interaction, which contributes to the stability of the immunological synapse via the formation of co-stimulatory B7–CD28 interactions between T-cells and B-cells (Klaus *et al.*, 1994). The ligation of CD40 stimulates the production in B-cells of cytokines, such as IL2, IL6, IL10, TNFα, LTα, LTβ and TGFβ (Clark and Shu, 1990; Burdin *et al.*, 1993; Kindler *et al.*, 1995; Worm and Geha, 1995; Worm *et al.*, 1998). The CD40–CD40L interaction also induces T-cells to produce cytokines that determine the antibody class to be expressed in B-cells, and contributes to the proliferation of B-cells (Finkelman *et al.*, 1990).

Signalling pathways activated by the ligation of CD40 originate from the interaction of the intracellular domain of the receptor with TNF-associated proteins (TRAFs) (Harigai *et al.*, 2004). As a consequence of the interaction of CD40 with its trimeric ligand, it forms clusters at the B-cell membrane. The clustering of the receptor involves the recruitment and localisation of the TRAFs to membrane microdomains, which enables them to initiate signalling cascades by interacting with downstream signalling proteins (Hostager *et al.*, 2000). Like many TNFR family members, CD40 activates

© The Author 2007. Published by Oxford University Press. All rights reserved.
For Permissions, please e-mail: journals.permissions@oxfordjournals.org

the JNK/SAPK and NF-κB pathways (Berberich *et al.*, 1994, 1996). Both pathways involve protein serine/threonine kinases that activate AP1 and Rel transcription factors, thereby regulating gene expression. The p38 kinase pathway, which leads to the activation of transcription factors such as ATF2 (Raingeaud *et al.*, 1996), has also been reported to be activated by CD40 (Sutherland *et al.*, 1996). The extracellular signal-regulated kinase/mitogen-activated protein kinase pathway, which is also activated by CD40 (Li *et al.*, 1996), contributes to the activation of AP1, NF-κB and NF-AT, and the subsequent induction of cytokine gene expression (Park and Levitt, 1993).

The mutation registry for XHIGM, CD40Lbase (Piirilä *et al.*, 2006; Notarangelo and Peitsch, 1996) (http://bioinf.uta.fi/CD40Lbase/), currently lists 212 XHIGM patient entries with a total of 128 different mutations. Most disease causing mutations are found in exons (106), 35 of which are missense mutations located mainly in the TNFH domain of the protein (Fig. 1A). We investigated the consequences of all the CD40L missense mutations by applying structural and bioinformatics methods.

Materials and methods

The amino acid sequence and missense mutations for CD40L were obtained from our database CD40Lbase (http://bioinf.uta.fi/CD40Lbase/). The database lists all known mutations for CD40L and stores patient information, such as clinical and immunological phenotype, prognosis and treatment. The database was updated with recently published cases. Sequence homologues (33) were obtained by PSI-BLAST (Altschul *et al.*, 1997), and homologues for the TNFH domain sequence were collected from the Pfam database (Bateman *et al.*, 2004) (seed: 52, full: 175 sequences). Multiple sequence alignments were performed by Clustal W (Thompson *et al.*, 1994). Alignments were visualised using MultiDisp (Riikonen, P. and Vihinen, M., in preparation) and ConSeq (Berezin *et al.*, 2004) for illustration of conserved amino acids in the sequence.

The evolutionary conservation of the sequences was studied, in addition to the visualisation programs, by ProCon, a program for calculating mutual information and entropy in amino acid sequences (Shen and Vihinen, 2004). For entropy calculations, default parameters ($p_1 = 0.01$, $p_2 = 0.05$) were used, whereas parameters for the mutual information were readjusted to $p_1 = 0.005$ and $p_2 = 0.020$. Conservation indices were calculated with the program al2co (Pei and Grishin, 2001) and the ConSurf server (Glaser *et al.*, 2003).

Structural disorder in the protein and the effects of mutations on disordered regions were studied using four predictors, DISOPRED (Ward *et al.*, 2004), DisEMBL (Linding *et al.*, 2003a), GlobPlot (Linding *et al.*, 2003b) and PONDR (Romero *et al.*, 1997). The disorder prediction methods are based on different principles, which are further discussed in the corresponding papers and in Thusberg and Vihinen (2006).

The effects of mutations on aggregation propensities were studied by TANGO (Fernandez-Escamilla *et al.*, 2004), and calculations presented by Chiti *et al.* (2003), for which α-helical propensities were calculated with the program AGADIR (Muñoz and Serrano, 1997). A script was written to implement the method of Chiti *et al.* (2003).

The damaging effects of point mutations were analysed using SNPs3D (Yue *et al.*, 2006), SIFT (Ng and Henikoff, 2001), PolyPhen (Sunyaev *et al.*, 2001), PoPMuSiC (Gilis and Rooman, 2000; Kwasigroch *et al.*, 2002) and Pmut (Ferrer-Costa *et al.*, 2005).

Structural analyses were performed based on the crystal structure of the protein (PDB 1ALY). The structure was visualised and the mutations were modelled by PyMOL (DeLano, 2002). Hydrogen atoms were added to the structures using Reduce (Word *et al.*, 1999b). Mutant amino acid side chain χ angles were rotated at intervals of 10° by the Autobondrot function in PROBE (Word *et al.*, 1999a; Lovell *et al.*, 2000) and the best rotamers were selected for further analysis. The acceptable conformations for a mutated side chain have a total score above -1.0, allowing for small local perturbations in the structure (Lovell *et al.*, 2000). The created structures were verified by MolProbity (Lovell *et al.*, 2003), which was also used for converting the PDB files into Kinemage format. MolProbity adds all atom contacts into the structures and flips asparagines and glutamine side chains when necessary. Mutation structures were visualised by the program KiNG (Lovell *et al.*, 2003), to analyse all atom contacts and clashes.

Amino acid contact analysis for the mutant residues in the TNFH domain was performed with CSU (Sobolev *et al.*, 1999), and the nature of the contacts, contact surfaces, as well as solvent accessible surfaces, were elucidated. Contact energies between amino acids in the TNFH domain were analysed using RankViaContact (Shen and Vihinen, 2003). By analysing the wild-type protein, we could determine structurally important amino acids, which contribute to the stability of the protein, or amino acids with weak contacts that may be important for functional specificity. The analysis of changes in the contact energies for mutant structures provided hypotheses for the roles of the mutated amino acids. Electrostatic surface potentials were calculated and visualised with the PyMOL program (DeLano, 2002) using the absolute electrostatic potential in a vacuum.

Results

Diseases can arise from numerous genetic defects. To understand the basis of diseases, one has to study the effects of mutations at the gene and/or protein level. Missense mutations are, in this respect, the most interesting because of the possibility to learn about functions and properties of the protein. The effects of many other mutation types are self evident, such as large deletions or insertions, frameshift mutations and nonsense mutations, which affect the size and/or sequence of the protein. All the reported missense mutations causing XHIGM, altogether 35, could be explained at the molecular level by means of sequence and structure analysis.

Sequence conservation and mutations at the conserved residues

Disease causing mutations are typically located at conserved positions within a protein family, since these positions are usually essential for the structure and/or function of the protein (Miller and Kumar, 2001; Mooney and Klein, 2002; Shen and Vihinen, 2004). In CD40L, the degree of conservation depends on the protein domain. The CD40L

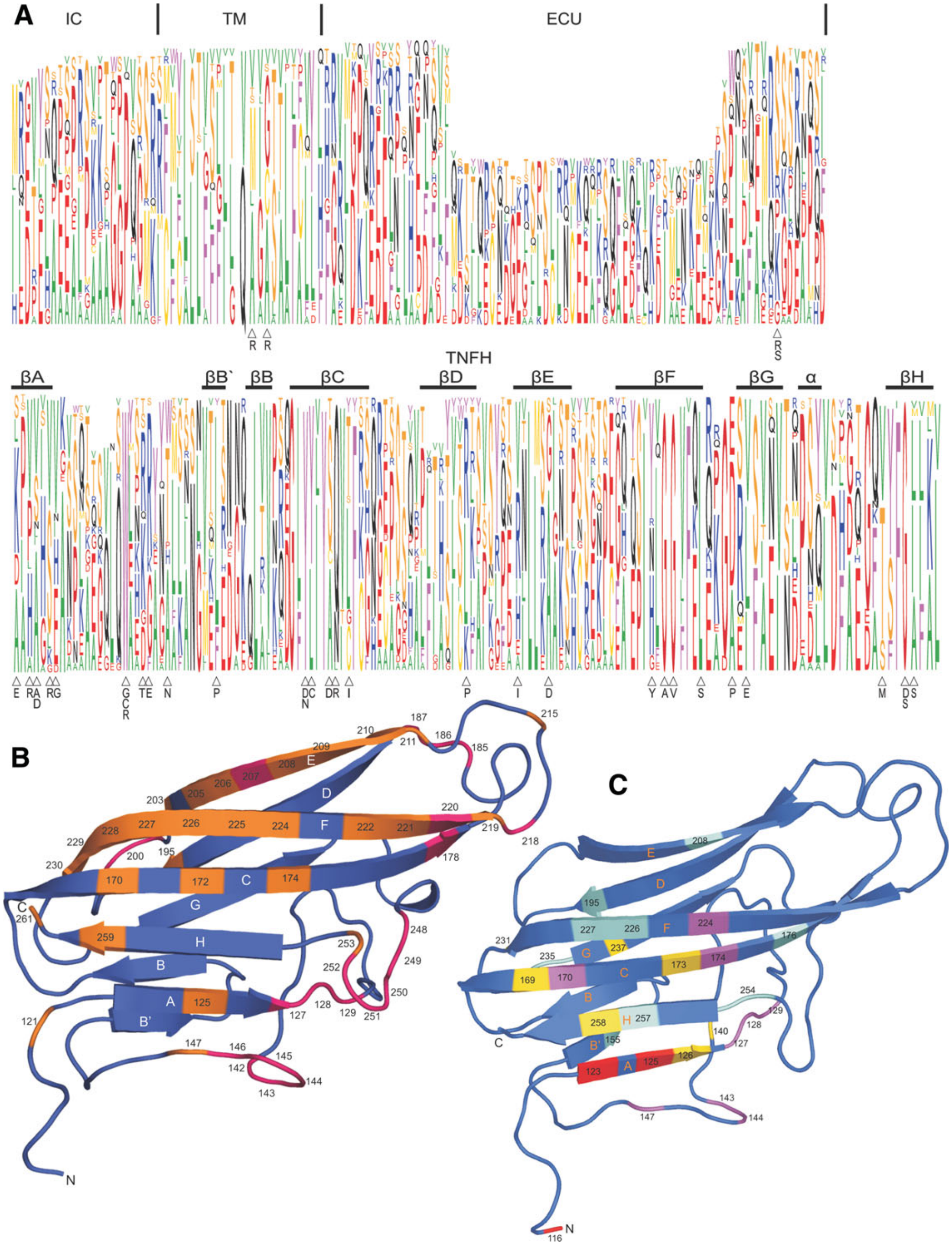


Fig. 1. (**A**) MultiDisp visualisation of the sequence alignment for CD40L and its homologues. The height of the characters indicates the frequency of the amino acids in the alignment positions, and the colour of the objects reflects the chemical nature of the amino acids. The CD40L domain boundaries and secondary structures are presented according to Karpusas *et al.* (1995). The positions of CD40L missense mutations are indicated by arrowheads below the alignment, together with all mutant forms. XHIGM-causing mutations are clustered almost exclusively to the TNFH domain. The protein consists of an intracellular domain, IC; a transmembrane domain, TM; an extracellular unique domain, ECU; and a TNF homology domain, TNFH. (**B**) Structure of CD40 ligand (PDB code 1ALY). The residues involved in protein–protein interactions are coded as follows: trimer formation – orange; receptor binding – magenta. (**C**) CD40L missense mutations coloured according to their principal effects on CD40L structure and function. Change in electrostatic surface potential – red; conformational perturbation – cyan; loss of hydrophobic interactions and structural stability – yellow; protein–protein interactions – magenta. Secondary structures are named according to Karpusas *et al.* (1995).

TNFH domain (residues 123–261) has many more homologues than the IC and ECU domains, which are very different from the corresponding domains of other family members.

According to the full sequence alignment visualised with MultiDisp (Riikonen *et al*, in preparation) (Fig. 1A) and the conservation indices calculated by al2co (Pei and Grishin, 2001) and ConSeq (Berezin *et al.*, 2004), there are

10 invariant positions in the TNF family corresponding to the amino acids W140, L161, G167, Y169, Y172, L205, G226, G227, L231 and G257 in CD40L. There are missense mutations at six of these positions: W140C/G/R, Y169D/N, G226A, G227V, L231S and G257D/S. Type II conservation, where the physicochemical nature of the amino acid is conserved, was studied by calculation of information with an alphabet in which amino acids are split into six groups based on their physicochemical properties. Type II conserved amino acids with known XHIGM causing mutations are Q174, with polarity as the conserved property, and Y170 and V237, the conserved property of which is hydrophobicity. In Fig. 1A, the physicochemical nature of these positions is represented by different colours.

Type III conservation refers to covariation of two or more positions in the protein family. In the TNF family, Type III conservation is evident, but almost all of the covarying amino acids are not conserved in the CD40L sequence—only at positions 222, 239 and 240 is the covarying amino acid the same as in the other family members, but there are no mutations in these amino acids.

Mutations predicted to affect structural disorder and protein β-aggregation propensity

None of the mutations was predicted to cause disorder by all the programs we used (DisEMBL (Linding *et al.*, 2003a), DISOPRED (Ward *et al.*, 2004), GlobPlot (Linding *et al.*, 2003b) and PONDR (Li *et al.*, 1999)). G116R is the only mutation likely to cause disorder, because three of the four programs agreed on the disorder-causing nature of the mutation. In addition, V126D, W140G, L155P, A208D, A235P, V237E and L258S might increase disorder in the protein structure, being predicted to do so by half of the programs. E129G, K143T, T176I, H224Y and G227V were predicted to increase the protein aggregation rate when calculated by the methods of Chiti *et al.* (2003). T176I was also predicted to cause aggregation by the program TANGO (Fernandez-Escamilla *et al.*, 2004; Linding *et al.*, 2004). In Fig. 1C, these effects are presented under the category of structural stability loss.

Structural mutations

In the structure-based studies, only the 33 mutations located within the structurally determined CD40L TNFH domain (PDB ID 1ALY) (Fig. 1A and C) could be analysed. At the position 116, where there are two known missense mutations, the structure is not well defined (Karpusas *et al.*, 1995). Consequently, the effects of these mutations cannot be reliably predicted at the structural level. Effects of mutations on protein structure and stability were studied by rotamer analysis and determination of overlapping side chains. The best rotamers according to the PROBE scores were used in the analyses. Most of the mutated side chains fit into the structure without deleterious changes to protein scaffolding, as determined both computationally by the PROBE score (Word *et al.*, 1999a; Word *et al.*, 2000) and visually by the program KiNG (Lovell *et al.*, 2003). Of the 33 mutations screened, 22 gave an acceptable score above −1.0, allowing for small local perturbations in the structure (Lovell *et al.*, 2000). In the visual inspection, 17 of the mutations showed very little or no effect on the structure.

According to PROBE scores, the S128R/E129G double mutant, L155P, T176I, L195P, G226A, G227V, A235P, T254M, G257D and G257S, do not fit to the structure in any rotamer (example in Fig. 2A). In addition, mutations A173D, W140C, W140R and A208D were shown to cause serious clashes with other side chains (example in Fig. 2B). Proline is a known secondary structure breaker, and L155 and A235 are located in the middle of β strands. Prolines in these positions have been suggested to cause structural disorder (Karpusas *et al.*, 1995). Mutated amino acids that cannot fit into the structure without clashes, lead to changes in protein scaffolding, stability and properties of the protein. These mutations are indicated as conformation perturbating in Fig. 1C.

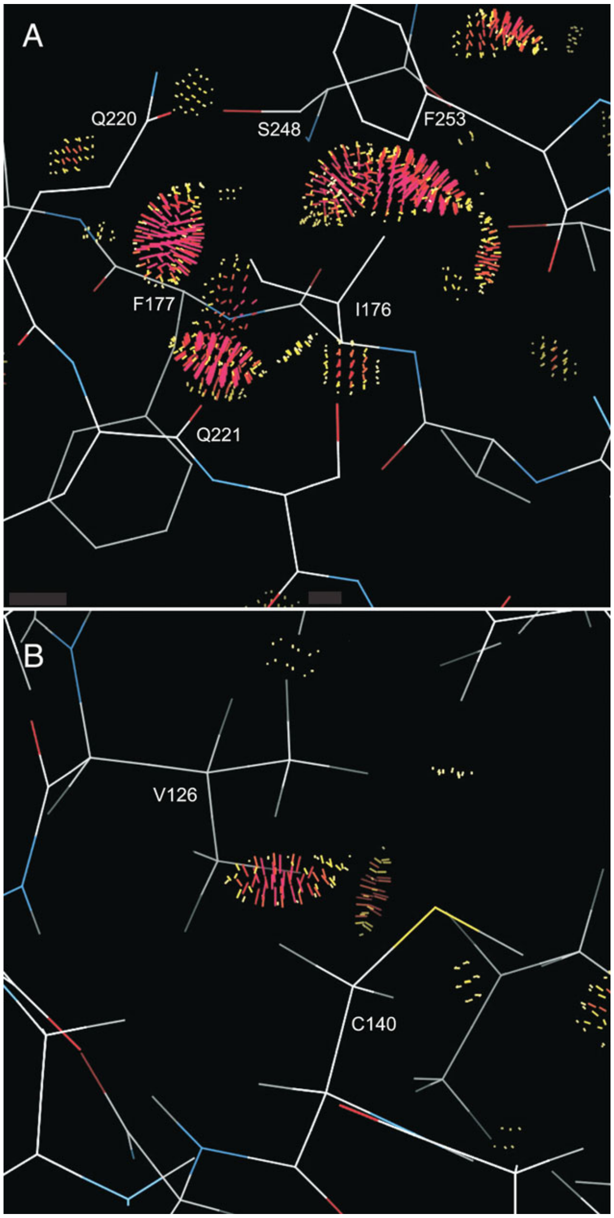


Fig. 2. (**A**) Substitution of T176 by I causes serious clashes with neigh bouring residues, also indicated by a negative PROBE score (−18.604). (**B**) Substitution of W140 by C causes serious clashes with V126, in spite of a positive PROBE score (0.639). Yellow – negligible overlap; red – significant overlap ≥0.25 Å; hot pink – serious clash overlap ≥0.4 Å.

Mutations causing changes in contacts maintaining stability

Amino acids located in the core of the protein, with a negligible solvent accessible surface area, typically form several hydrophobic interactions essential for the folding of the protein and for the stability of the protein structure. Mutations in such amino acids may cause detrimental changes to the structure-maintaining contacts. The probability of a mutation being pathogenic has been shown to increase with a decrease in the solvent accessibility of the site (Vitkup *et al.*, 2003). In addition, introduction of charged side chains into the hydrophobic core is known to destabilise protein structure (Chasman and Adams, 2001). Of the 33 CD40L TNFH mutations, 10 cause significant loss of hydrophobic interactions: V126A, V126D, W140G, W140R, W140C, Y169D, Y169 N, A173D, L231S and L258S. All these residues are located in the hydrophobic core of the protein, with a solvent accessible surface of 0.0–1.4%, and strong contact energies.

Residues with strong contact energies are important for protein stability (Shen and Vihinen, 2003). Mutations that affect such residues can thus be predicted to decrease stability. Of the 33 investigated mutations, five affect residues from among the 10% most stabilising amino acids: V126D, V237E, G257D, G257S and L258S. V237 forms several hydrophobic interactions with neighbouring residues, but the mutation does not change the number of these contacts. Instead, the contact energy of the residue decreases from −23.850 to −5.550, which indicates a significant weakening of the contacts as a consequence of the mutation. G257, another residue with strong contact energy, is also presumably structurally important. The hydrogen bonds it forms with Y172 and L258 do not change upon substitution with D or S, so the structure destabilising effects of the mutations can be explained by side chain clashes and the restriction of the mobility of the backbone (as usually happens when glycine is replaced by an alternative amino acid). The introduction of a charged residue (aspartate) into the protein core would require another mutation to neutralise the charge.

L258 and V126 form several strong hydrophobic contacts, the number of which decreases markedly when mutated to an S or D. The negative charge introduced by aspartate into the hydrophobic core contributes to the destabilising effect of the mutation V126D. The mutations, whose principal effect is the loss of essential stability maintaining contacts, are categorised as stability reducing in Fig. 1C.

Mutations affecting the electrostatic surface potential of CD40L

G116R, A123E, H125R, V126D, S128R/E129G, K143T, G144E, A173D, Q174R, R203I, A208D, V237E and G257D introduce significant changes into the electrostatic surface potential of CD40L (Figs. 1C and 3). Most mutations change the surface potential more negative, whereas the arginine-introducing mutations, G116R, H125R, Q174R and S128R/E129G, make the potential more positive. W140R does not introduce a significant change to the electrostatic surface potential because the side chain of the mutated residue lies inside the structure. However, the positive charge is not

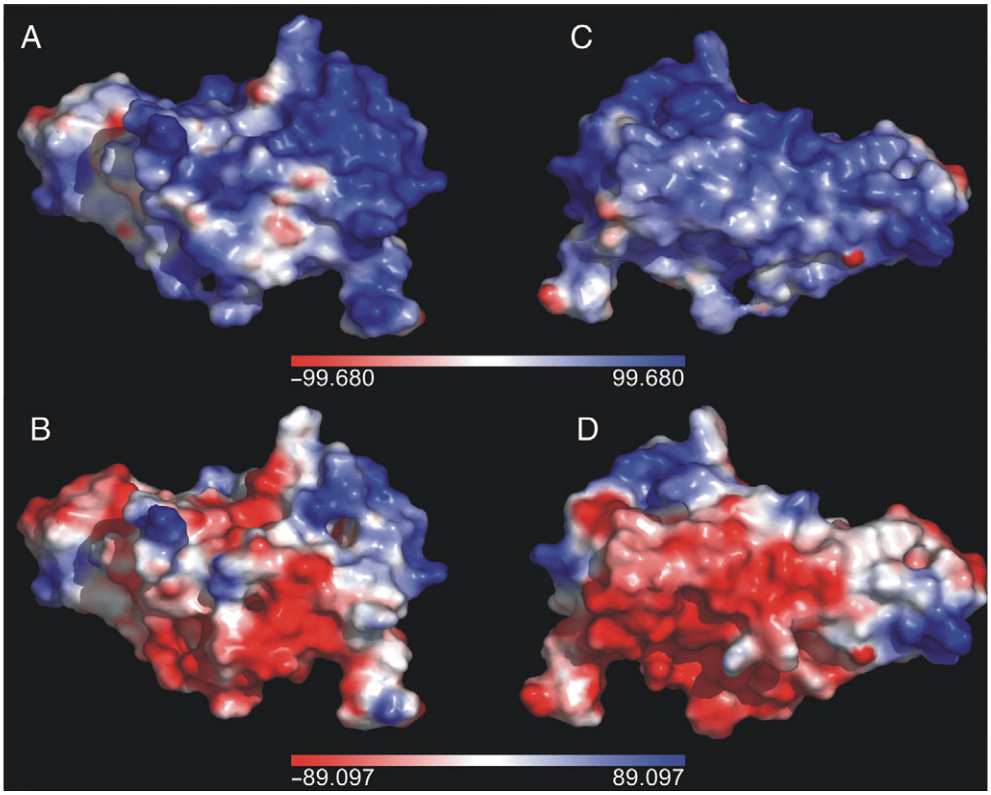


Fig. 3. Electrostatic surface potentials in wild-type (**A** and **C**) and mutated (**B** and **D**) CD40L. In the mutated structures, all missense mutations are included, except for the cases where a single position is affected by several mutations. In these positions, the following mutations are displayed: G116R, V126A, W140R, Y169 N and G257D. A large proportion of the mutations alter the surface potential from positive to negative.

neutralised by interactions with other residues in the protein core.

Changes in the electrostatic potential affect the properties of a protein in many ways. Electrostatics is a significant factor in protein folding and stability, and it has an effect on protein interactions. Electrostatic surface potential has a major role in CD40 ligand–receptor interactions, since the positive surface of the ligand attracts the negative surface of the receptor (Singh *et al.*, 1998).

Effects of mutations on protein–protein interactions

The function of the CD40 ligand is based on two different kinds of protein–protein interactions. The interaction with the receptor is essential for initiating the signalling cascade leading to B-cell activation, and interactions between CD40L monomers enable CD40L trimer formation, which has to occur in order for the interaction with the receptor to take place (Peitsch and Jongeneel, 1993).

The CD40L–CD40 interaction is based on electrostatic interactions whereby basic side chains on the CD40L surface attract acidic side chains on the surface of CD40. CD40L residues K143, R203 and R207 form salt bridges with receptor surface amino acids (Singh *et al.*, 1998). K143 is mutated to T in XHIGM, and thus a salt bridge with the receptor E66 is lost. K143T also affects the CD40L electrostatic surface potential, which contributes to the loss of affinity between the ligand and receptor. The substitution of R203 by I leads to the loss of the ion pair formed with E74 of the CD40. The mutation changes the surface potential more negative as well.

In addition to the acid–base contacts formed between the ligand–receptor pair, other direct contact (distance between heavy atoms <5 Å) forming CD40L residues are: I127, S128, E129, E142, G144, Y145, Y146, C178, S185, Q186, A187, R200, F201, C218, Q220, S248, H249, G250, T251 and G252 (Singh *et al.*, 1998). XHIGM causing mutations occur in residues 128, 129 and 144. When a glycine is replaced by a glutamate; with a long side chain and a negative charge, it has an effect on the specificity of the interaction. The introduction of glutamate might result in the loss of the conformational freedom necessary at that position in order to form the corner of the AA″ loop (Karpusas *et al.*, 1995). The orientation of K143 and Y145 might consequently change, making the ligand–receptor interaction less likely to occur.

The essential contact-forming residues in homotrimer interactions are Y170 and H224 (Karpusas *et al.*, 1995), both of which are mutated in XHIGM (Y170C and H224Y). Neither of these mutations changes the polar nature of the position, but specific contacts between the monomers are potentially affected or hindered. Complex stabilising interactions also important for CD40L trimerisation are formed by Q121, H125, Y145, T147, Y172, Q174, L195, R203, L205, L206, R207, A208, A209, N210, T211, A215, G219, Q220, Q221, S222, L225, G226, G227, V228, F229, E230, T251, G252, F253, L259 and L261 (Morris *et al.*, 1999; Karpusas *et al.*, 2001). There are eight mutations at positions 125, 147, 174, 195, 203, 208, 226 and 227. The residues involved in protein–protein interactions are presented in Fig. 1B. The effects of mutations on the structure and function of CD40L are summarised in Fig. 1C and Table I.

Discussion

Here, we have investigated the structural effects and consequences of disease-causing mutations in CD40 ligand. We have collected information about disease-causing mutations in immunodeficiencies to databases called IDbases (Piirilä *et al.*, 2006). Currently there are 115 IDbases and 4587 patient cases in them. We have previously applied bioinformatics and structural analysis methods to reveal the basis of e.g. Bruton's tyrosine kinase mutations in X-linked agammaglobulinemia (Vihinen *et al.*, 1994a, 1994b; Väliaho *et al.*, 2006), SHD1A mutations in X-linked lymphoproliferative disease (Lappalainen *et al.*, 2000), BLM mutations in Bloom syndrome (Rong *et al.*, 2000), mutations in the WAS protein in Wiskott–Aldrich syndrome (Rong and Vihinen, 2000) and mutations in the methyltransferase domain of DNMT3B in immunodeficiency, centromeric instability and facial abnormalities (ICF) syndrome (Lappalainen and Vihinen, 2002). In addition, we have tested and discussed the applicability of sequence and structure-based bioinformatics methods to reveal structure–function correlations of disease-causing missense mutations (Thusberg and Vihinen, 2006). In addition to understanding the molecular basis of disease, the ability to predict the effects of amino acid substitutions is useful for protein engineering purposes.

Most disease causing mutations affect the stability of protein structure (Wang and Moult, 2001; Steward *et al.*, 2003). Thirteen of the thirty-five mutations (one being a double mutant) in CD40L can be classified as functional, directly changing amino acids involved in trimerisation and ligand–receptor interactions. Eight of these mutated amino acids are also involved in stabilisation of the CD40L trimer, which is why they could also be classified as structural mutations (Table I). Because of the correlation between structure and function, the classifications are overlapping.

Conserved amino acids tend to be essential for structure and function, which is why disease-causing mutations often occur at the corresponding positions (Miller and Kumar, 2001; Mooney and Klein, 2002). The probability that a random mutation will cause a genetic disease has been shown to increase with an increase in the degree of site conservation (Vitkup *et al.*, 2003). In the TNF family, sequence conservation is evident, and 37% of CD40L mutations affect Type I and Type II conserved amino acids (Table I, Fig. 1A). There are many covarying positions in the protein family, but only few of them are conserved in CD40L. None of these sites has been shown to have disease-causing mutations.

The members of the TNF family exhibit structural conservation—all of them have a jelly roll fold and Greek key topology. The specific functions of these proteins are governed by the loops connecting the β strands. The length and properties of the loops vary significantly among the family members (Karpusas *et al.*, 1995). Most of the disease-causing mutations in CD40L are located in the β strands (Fig. 1A and C), and are thus predicted to affect protein structure and stability, thereby hindering protein function.

Some missense mutations may increase disorder in the CD40L structure according to our predictions. Although there are several methods available for disorder prediction, they seldom agree on the effects of mutations. The structure-based predictions of the effects of mutations gave further

Table I. Summary of the effects of CD40L mutations on structure and function

Mutation	Conserved residue	Electrostatic surface potential	Contacts and stability	Disorder	Conformational	Aggregation	Protein–protein interactions	Other
M36R								Lowered expression on the membrane[a]
G38R								Lowered expression on the membrane[a]
G116R		X		X				Structure not well defined[b]
G116S								Structure not well defined[b]
A123E		X						
H125R		X					X	
V126A			X					
V126D		X	X	X				Introduction of a charged residue into the hydrophobic core
S128R/E129G		X			X	X	X	Loss of conformational stability[b]
W140G	X		X	X				Loss of conformational stability[b]
W140C	X		X		X			
W140R	X		X		X			Loss of conformational stability.[b] Introduction of a charged residue into the hydrophobic core
K143T		X				X	X	
G144E		X					X	Loss of conformational freedom.[b] Possible change in the orientation of residues 143 and 145 involved in receptor contact
T147 N							X	
L155P				X	X			Disruption of a β strand[b]
Y169D	X		X					Introduction of a charged residue into the hydrophobic core
Y169 N	X		X					
Y170C	X						X	
A173D		X	X		X			Introduction of a charged residue into the hydrophobic core
Q174R	X	X					X	
T176I					X	X		
L195P					X		X	
R203I		X					X	
A208D		X		X	X		X	Introduction of a charged residue into the hydrophobic core
H224Y						X	X	
G226A	X				X		X	
G227V	X				X	X	X	
L231S	X		X					
A235P				X	X			Disruption of a β strand[b]
V237E	X	X	X	X				Contacts are preserved but weakened. Introduction of a charged residue into the hydrophobic core
T254M					X	X		
G257D	X	X	X		X			Contacts are preserved but weakened. Restriction of backbone mobility. Introduction of a charged residue into the hydrophobic core
G257S	X		X		X	X		Contacts are preserved but weakened. Restriction of backbone mobility
L258S			X	X				

[a] Garber *et al.* (1999).
[b] Karpusas *et al.* (1995).

insight into their role in CD40L structure and function. The three-dimensional structure has been determined only for the TNFH domain, thus missense mutations outside the region (M36R and G38R) could not be analysed at the structural level. At the first residues of the structurally determined domain, the structure is not well defined (Karpusas *et al.*, 1995), which is why the effects of the mutations G116R and G116S cannot be reliably predicted at the structural level. The mutations are likely to cause conformational rearrangements into the structure. Mutations introducing arginine can be predicted to affect protein structure, as problems in side chain packing are common when the replacing residue is larger than the one being substituted, especially when the substituted amino acid is glycine. A positively charged side chain may also change fundamental structure-maintaining contacts. In the membrane spanning α-helix, the introduction of positively charged amino acids may cause problems for the stability of the helix or its insertion into the membrane. It has been hypothesised that positively charged substitutions in transmembrane helices act as signals guiding the protein to be degraded in the endoplasmic reticulum (Bonifacino *et al.*, 1991). Indeed, CD40L forms with the mutations M36R or G38R are expressed on the T-cell membrane to a greatly reduced extent (10%) compared to the wild-type protein (Garber *et al.*, 1999) (Table I). The mutations in the transmembrane helix probably cause the XHIGM phenotype by reducing the expression of CD40L on the T-cell surface, thereby decreasing the number of ligand–receptor contacts.

Forty percent of the missense mutations analysed cause major structural changes to the protein, according to the PROBE score for the best rotamer (Table I). Side chains with a low score do not fit into the structure in any conformation, causing changes to the structure already during the folding process. Side chains predicted to cause clashes lead to at least local rearrangements of the structure as well. Even subtle changes in protein scaffolding may have an influence on specific protein–protein interactions. The amino acid contacts in the hydrophobic core are crucial for the folding and stability of the protein. Thirteen of the mutations in CD40L affect amino acids forming strong contacts in the hydrophobic core of the protein, thereby causing the loss of a number of structure and stability maintaining contacts (Table I).

Electrostatic surface potentials, calculated with the program PyMOL, are suggestive and qualitative (DeLano, 2002). Electrostatic surface potential is an important property of CD40L, since the ligand–receptor interaction is mainly based on the attraction of molecular surfaces having opposite charges (Singh *et al.*, 1998). Our results indicate that the changes in the potential were evident for many of the substitutions (Fig. 3, Table I). Mutations that make the potential more negative are likely to affect the affinity between the ligand and receptor, because the positive surface potential of CD40L attracts the negative surface of CD40 (Singh *et al.*, 1998).

The consequences of mutations are diverse and the different effects on CD40L structure and function are equally represented. Thirty-seven percent of the mutations affect residues known to be crucial for receptor binding or trimerisation (Table I). Electrostatic surface potential, which is also an important factor in protein–protein interactions, is affected by six additional substitutions (Table I). Thus, more than 50% of XHIGM causing missense mutations are predicted to affect CD40L ligation and trimerisation (Table I), the proportion of structural mutations being slightly bigger (63%). Generally, the majority of pathogenic mutations affect structural rather than functional residues (Wang and Moult, 2001; Mooney and Klein, 2002). The analysis of structural and functional consequences of the CD40L mutations identified in XHIGM patients provides insights into the molecular basis of the syndrome. Further analysis at the experimental level will be needed to test our predictive findings and to fully understand the mechanisms behind the disease.

Acknowledgements

The financial support from the Tampere Graduate School in Biomedicine and Biotechnology, the Medical Research Fund of the Tampere University Hospital and the EU is gratefully acknowledged.

References

Allen,R.C. *et al.* (1993) *Science*, **259**, 990–993.

Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) *Nucleic Acids Res.*, **25**, 3389–3402.

Armitage,R.J. *et al.* (1992) *Nature*, **357**, 80–82.

Aruffo,A. *et al.* (1993) *Cell*, **72**, 291–300.

Bateman,A. *et al.* (2004) *Nucleic Acids Res.*, **32**, D138–D141.

Berberich,I., Shu,G., Siebelt,F., Woodgett,J.R., Kyriakis,J.M. and Clark,E.A. (1996) *EMBO J.*, **15**, 92–101.

Berberich,I., Shu,G.L. and Clark,E.A. (1994) *J. Immunol.*, **153**, 4357–4366.

Berezin,C., Glaser,F., Rosenberg,J., Paz,I., Pupko,T., Fariselli,P., Casadio,R. and Ben-Tal,N. (2004) *Bioinformatics*, **20**, 1322–1324.

Bodmer,J.L., Schneider,P. and Tschopp,J. (2002) *Trends Biochem. Sci.*, **27**, 19–26.

Bonifacino,J.S., Cosson,P., Shah,N. and Klausner,R.D. (1991) *EMBO J.*, **10**, 2783–2793.

Burdin,N., Peronne,C., Banchereau,J. and Rousset,F. (1993) *J. Exp. Med.*, **177**, 295–304.

Chasman,D. and Adams,R.M. (2001) *J. Mol. Biol.*, **307**, 683–706.

Chiti,F., Stefani,M., Taddei,N., Ramponi,G. and Dobson,C.M. (2003) *Nature*, **424**, 805–808.

Clark,E.A. and Ledbetter,J.A. (1986) *Proc. Natl Acad. Sci. USA*, **83**, 4494–4498.

Clark,E.A. and Shu,G. (1990) *J. Immunol.*, **145**, 1400–1406.

Coffman,R.L., Lebman,D.A. and Rothman,P. (1993) *Adv. Immunol.*, **54**, 229–270.

Dadgostar,H., Zarnegar,B., Hoffmann,A., Qin,X.F., Truong,U., Rao,G., Baltimore,D. and Cheng,G. (2002) *Proc. Natl Acad. Sci. USA*, **99**, 1497–1502.

DeLano,W.L. (2002) DeLano Scientific, San Carlos, CA, USA. http://www.pymol.org

Fernandez-Escamilla,A.M., Rousseau,F., Schymkowitz,J. and Serrano,L. (2004) *Nat. Biotechnol.*, **22**, 1302–1306.

Ferrer-Costa,C., Gelpi,J.L., Zamakola,L., Parraga,I., de la Cruz,X. and Orozco,M. (2005) *Bioinformatics*, **21**, 3176–3178.

Finkelman,F.D., Holmes,J., Katona,I.M., Urban,J.F., Beckmann,M.P., Park,L.S., Schooley,K.A., Coffman,R.L., Mosmann,T.R. and Paul,W.E. (1990) *Annu. Rev. Immunol.*, **8**, 303–333.

Fuleihan,R., Ramesh,N., Loh,R., Jabara,H., Rosen,R.S., Chatila,T., Fu,S.M., Stamenkovic,I. and Geha,R.S. (1993) *Proc. Natl Acad. Sci. USA*, **90**, 2170–2173.

Garber,E., Su,L., Ehrenfels,B., Karpusas,M. and Hsu,Y.M. (1999) *J. Biol. Chem.*, **274**, 33545–33550.

Gilis,D. and Rooman,M. (2000) *Protein Eng.*, **13**, 849–856.

Glaser,F., Pupko,T., Paz,I., Bell,R.E., Bechor-Shental,D., Martz,E. and Ben-Tal,N. (2003) *Bioinformatics*, **19**, 163–164.

Harigai,M. *et al.* (2004) *Arthritis Rheum.*, **50**, 2167–2177.

Hostager,B.S., Catlett,I.M. and Bishop,G.A. (2000) *J. Biol. Chem.*, **275**, 15392–15398.

Karpusas,M., Hsu,Y.M., Wang,J.H., Thompson,J., Lederman,S., Chess,L. and Thomas,D. (1995) *Structure*, **3**, 1031–1039.

Karpusas,M., Lucci,J., Ferrant,J., Benjamin,C., Taylor,F.R., Strauch,K., Garber,E. and Hsu,Y.M. (2001) *Structure*, **9**, 321–329.

Kindler,V., Matthes,T., Jeannin,P. and Zubler,R.H. (1995) *Eur. J. Immunol.*, **25**, 1239–1243.

Klaus,S.J., Berberich,I., Shu,G. and Clark,E.A. (1994) *Semin. Immunol.*, **6**, 279–286.

Kroczek,R.A., Graf,D., Brugnoni,D., Giliani,S., Korthuer,U., Ugazio,A., Senger,G., Mages,H.W., Villa,A. and Notarangelo,L.D. (1994) *Immunol. Rev.*, **138**, 39–59.

Kwasigroch,J.M., Gilis,D., Dehouck,Y. and Rooman,M. (2002) *Bioinformatics*, **18**, 1701–1702.

Lappalainen,I., Giliani,S., Franceschini,R., Bonnefoy,J.Y., Duckett,C., Notarangelo,L.D. and Vihinen,M. (2000) *Biochem. Biophys. Res. Commun.*, **269**, 124–130.

Lappalainen,I. and Vihinen,M. (2002) *Protein Eng.*, **15**, 1005–1014.

Levy,J. *et al.* (1997) *J. Pediatr.*, **131**, 47–54.

Li,X., Romero,P., Rani,M., Dunker,A.K. and Obradovic,Z. (1999) *Genome Inform. Ser. Workshop Genome Inform.*, **10**, 30–40.

Li,Y.Y., Baccam,M., Waters,S.B., Pessin,J.E., Bishop,G.A. and Koretzky,G.A. (1996) *J. Immunol.*, **157**, 1440–1447.

Linding,R., Jensen,L.J., Diella,F., Bork,P., Gibson,T.J. and Russell,R.B. (2003a) *Structure*, **11**, 1453–1459.

Linding,R., Russell,R.B., Neduva,V. and Gibson,T.J. (2003b) *Nucleic Acids Res.*, **31**, 3701–3708.

Linding,R., Schymkowitz,J., Rousseau,F., Diella,F. and Serrano,L. (2004) *J. Mol. Biol.*, **342**, 345–353.

Lovell,S.C., Davis,I.W., Arendall,W.B., 3rd, de Bakker,P.I., Word,J.M., Prisant,M.G., Richardson,J.S. and Richardson,D.C. (2003) *Proteins*, **50**, 437–450.

Lovell,S.C., Word,J.M., Richardson,J.S. and Richardson,D.C. (2000) *Proteins*, **40**, 389–408.

Miller,M.P. and Kumar,S. (2001) *Hum. Mol. Genet.*, **10**, 2319–2328.

Mooney,S.D. and Klein,T.E. (2002) *BMC Bioinformatics*, **3**, 24.

Morris,A.E., Remmele,R.L. Jr, Klinke,R., Macduff,B.M., Fanslow,W.C. and Armitage,R.J. (1999) *J. Biol. Chem.*, **274**, 418–423.

Muñoz,V. and Serrano,L. (1997) *Biopolymers*, **41**, 495–509.

Ng,P.C. and Henikoff,S. (2001) *Genome Res.*, **11**, 863–874.

Noelle,R.J., Roy,M., Shepherd,D.M., Stamenkovic,I., Ledbetter,J.A. and Aruffo,A. (1992) *Proc. Natl Acad. Sci. USA*, **89**, 6550–6554.

Notarangelo,L.D., Duse,M. and Ugazio,A.G. (1992) *Immunodefic. Rev.*, **3**, 101–121.

Notarangelo,L.D. and Peitsch,M.C. (1996) *Immunol. Today*, **17**, 511–516.

Park,J.H. and Levitt,L. (1993) *Blood*, **82**, 2470–2477.

Pei,J. and Grishin,N.V. (2001) *Bioinformatics*, **17**, 700–712.

Peitsch,M.C. and Jongeneel,C.V. (1993) *Int. Immunol.*, **5**, 233–238.

Piirilä,H., Väliaho,J. and Vihinen,M. (2006) *Hum. Mutat.*, **27**, 1200–1208.

Pinchuk,L.M., Klaus,S.J., Magaletti,D.M., Pinchuk,G.V., Norsen,J.P. and Clark,E.A. (1996) *J. Immunol.*, **157**, 4363–4370.

Raingeaud,J., Whitmarsh,A.J., Barrett,T., Derijard,B. and Davis,R.J. (1996) *Mol. Cell Biol.*, **16**, 1247–1255.

Romero,P., Obradovic,Z. and Dunker,A.K. (1997) *Genome Inform. Ser. Workshop Genome Inform.*, **8**, 110–124.

Rong,S.B., Väliaho,J. and Vihinen,M. (2000) *Mol. Med.*, **6**, 155–164.

Rong,S.B. and Vihinen,M. (2000) *J. Mol. Med.*, **78**, 530–537.

Shen,B. and Vihinen,M. (2003) *Bioinformatics*, **19**, 2161–2162.

Shen,B. and Vihinen,M. (2004) *Protein Eng. Des. Sel.*, **17**, 267–276.

Singh,J., Garber,E., Van Vlijmen,H., Karpusas,M., Hsu,Y.M., Zheng,Z., Naismith,J.H. and Thomas,D. (1998) *Protein Sci.*, **7**, 1124–1135.

Sobolev,V., Sorokine,A., Prilusky,J., Abola,E.E. and Edelman,M. (1999) *Bioinformatics*, **15**, 327–332.

Steward,R.E., MacArthur,M.W., Laskowski,R.A. and Thornton,J.M. (2003) *Trends Genet.* **19**, 505–513.

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

Sutherland,C.L., Heath,A.W., Pelech,S.L., Young,P.R. and Gold,M.R. (1996) *J. Immunol.*, **157**, 3381–3390.

Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) *Nucleic Acids Res.*, **22**, 4673–4680.

Thusberg,J. and Vihinen,M. (2006) *Hum. Mutat.*, **27**, 1230–1243.

Väliaho,J., Smith,C.I.E. and Vihinen,M. (2006) *Hum. Mutat.*, **27**, 1209–1217.

Vihinen,M., Nilsson,L. and Smith,C.I.E. (1994a) *Biochem. Biophys. Res. Commun.*, **205**, 1270–1277.

Vihinen,M. *et al.* (1994b) *Proc. Natl Acad. Sci. USA*, **91**, 12803–12807.

Vitkup,D., Sander,C. and Church,G.M. (2003) *Genome Biol.*, **4**, R72.

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

Ward,J.J., McGuffin,L.J., Bryson,K., Buxton,B.F. and Jones,D.T. (2004) *Bioinformatics*, **20**, 2138–2139.

Word,J.M., Bateman,R.C., Jr, Presley,B.K., Lovell,S.C. and Richardson,D.C. (2000) *Protein Sci.*, **9**, 2251–2259.

Word,J.M., Lovell,S.C., LaBean,T.H., Taylor,H.C., Zalis,M.E., Presley,B.K., Richardson,J.S. and Richardson,D.C. (1999a) *J. Mol. Biol.*, **285**, 1711–1733.

Word,J.M., Lovell,S.C., Richardson,J.S. and Richardson,D.C. (1999b) *J. Mol. Biol.*, **285**, 1735–1747.

Worm,M., Ebermayer,K. and Henz,B. (1998) *Immunology*, **94**, 395–402.

Worm,M. and Geha,R.S. (1995) *Int. Arch. Allergy Immunol.*, **107**, 368–369.

Yue,P., Melamud,E. and Moult,J. (2006) *BMC Bioinform.*, **7**, 166.

Received September 19, 2006; revised November 17, 2006; accepted December 19, 2006

Edited by Elizabeth Meiering

PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

Genome wide analysis of pathogenic SH2 domain mutations

Ilkka Lappalainen,[1,2] Janita Thusberg,[2] Bairong Shen,[2] and Mauno Vihinen[2,3]*

[1] Department of Biological and Environmental Sciences, Division of Biochemistry, FI-00014 University of Helsinki, Finland
[2] Institute of Medical Technology, FI-33014 University of Tampere, Finland
[3] Research Unit, Tampere University Hospital, FI-33520 Tampere, Finland

ABSTRACT

The authors have made a genome-wide analysis of mutations in Src homology 2 (SH2) domains associated with human disease. Disease-causing mutations have been detected in the SH2 domains of cytoplasmic signaling proteins Bruton tyrosine kinase (BTK), SH2D1A, Ras GTPase activating protein (RasGAP), ZAP-70, SHP-2, STAT1, STAT5B, and the p85α subunit of the PIP3. Mutations in the BTK, SH2D1A, ZAP70, STAT1, and STAT5B genes have been shown to cause diverse immunodeficiencies, whereas the mutations in RASA1 and PIK3R1 genes lead to basal carcinoma and diabetes, respectively. PTPN11 mutations cause Noonan sydrome and different types of cancer, depending mainly on whether the mutation is inherited or sporadic. We collected and analyzed all known pathogenic mutations affecting human SH2 domains by bioinformatics methods. Among the investigated protein properties are sequence conservation and covariance, structural stability, side chain rotamers, packing effects, surface electrostatics, hydrogen bond formation, accessible surface area, salt bridges, and residue contacts. The majority of the mutations affect positions essential for phosphotyrosine ligand binding and specificity. The structural basis of the SH2 domain diseases was elucidated based on the bioinformatic analysis.

Proteins 2008; 72:779–792.
© 2008 Wiley-Liss, Inc.

Key words: SH2 domains; mutations; bioinformatics; mutation analysis; mutation database; structural basis of disease.

INTRODUCTION

Src homology 2 (SH2) domains are about 100 residues in length sharing on average 28% pairwise sequence identity. These domains associate almost invariably to phosphorylated tyrosine residues in specific sequence contexts, and thus specifically function in protein tyrosine kinase (PTK) pathways. The binding of SH2 domains to their targets recruits the SH2 domain-containing protein to its proper signaling complex and thereby initiates or regulates downstream signaling cascades. In addition to their role in assembling activated complexes, particular SH2 domains can also form intramolecular interactions that regulate enzyme activity.[1]

More than 100 SH2 domains have been found among the human gene products. They usually appear in multidomain proteins, together with catalytic domains, or other protein binding modules, such as Src homology 3 (SH3), phosphotyrosine binding (PTB), or pleckstrin homology (PH) domains.[2] Many SH2 domains have coevolved and multiplied simultaneously with SH3 and kinase domains in protein tyrosine kinases, which highlights the fact that these domains usually function in a concerted way.[3]

Genome wide search for mutations in the SH2 domains revealed eight proteins, Bruton tyrosine kinase (BTK), SH2 domain-containing protein 1A (SH2D1A), Ras GTPase activating protein (RasGAP), protein tyrosine kinase ZAP70 (ZAP-70), SHP-2, signal transducer and activator of transcription 1α/β (STAT1), STAT5B, and the p85α subunit of the phosphatidylinositol 3-kinase (PI3-K), which have been shown to cause 10 distinct clinical phenotypes. The affected genes, their protein products, and the diseases caused by the mutations are summarized in Table I. All the affected SH2 domains either have a crucial role during cell development process or regulate multiple signaling cascades.[4–9] The biological processes and partners of the disease-related SH2 domain-containing proteins are discussed in detail in the Supplementary text. Disease-causing human SH2 domain mutation types range from large gross deletions of the whole gene to single point mutations. Missense mutations account for more than half of all mutation types (Table II).

The Supplementary Material referred to in this article can be found online at http://www.interscience.wiley.com/jpages/0887-3585/suppmat/

Grant sponsors: Tampere Graduate School in Biomedicine and Biotechnology; Medical Research Fund of the Tampere University Hospital; Sigrid Juselius Foundation; Academy of Finland.

Ilkka Lappalainen and Janita Thusberg contributed equally to this work.

Ilkka Lappalainen's current address is European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, United Kingdom.
*Correspondence to: Mauno Vihinen, Institute of Medical Technology, FI-33014 University of Tampere, Finland. E-mail: mauno.vihinen@uta.fi

Received 9 July 2007; Revised 23 November 2007; Accepted 17 December 2007
Published online 7 February 2008 in Wiley InterScience (www.interscience.wiley.com).
DOI: 10.1002/prot.21970

Table I
Diseases Related to SH2 Domains

Affected gene	Protein	Disease	OMIM	Inheritance	Prevalence	Phenotypes
BTK	BTK	X-linked agammaglobulinemia (XLA)	300,300	X-linked	1:200,000 in males	Hypogammaglobulinemia, antibody deficiency, recurrent bacterial infections
SH2D1A	SH2D1A (SAP)	X-linked lymphoproliferative disease (XLP)	308,240	X-linked	1:1,000,000 in males	Fatal infectious mononucleosis, malignant B cell lymphomas, dysgammaglobulinemia
ZAP70	ZAP-70	Severe combined immunodeficiency (SCID)	600,802	Autosomal recessive	n.a.[a]	Severe pulmonary infection, chronic diarrhea, failure to thrive, persistent candidiasis
PTPN11	SHP-2	Noonan syndrome	163,950	Autosomal dominant	1:1000–1:2500	Short stature, facial dysmorphia, wide spectrum of congenital heart defects
PTPN11	SHP-2	Noonan-like/multiple giant-cell lesion syndrome	163,955	Autosomal dominant	n.a.	In addition to main Noonan syndrome phenotypes, giant-cell lesions of bone and soft tissues
PTPN11	SHP-2	Juvenile myelomonocytic leukaemia (JMML)	607,785	n.a.	n.a.	~30% of myelodysplastic syndrome and 2% of leukaemia patients
PIK3R1	P85α	Severe insulin deficiency	—	n.a.	n.a.	Acanthosis nigricans, hyperinsulinemia, diabetes mellitus at the later stage
RASA1	RasGAP	Basal cell carcinoma (BCC)	605,462	Sporadic	n.a.	Clusters of basal cell carcinoma, development of tumours on the chest
STAT1	STAT1	STAT1 deficiency, complete	600,555	n.a.	n.a.	Susceptibility to viral and intracellular bacterial infections
STAT5B	STAT5B	Growth hormone insensitivity with immunodeficiency	245,590	n.a.	n.a.	Growth failure, recurrent bacterial and viral infections

[a] n.a., not available.

Mutations can prevent the function of a protein in many ways. Therefore, attempts have been made to distinguish functional mutations from structural ones in SH2 domains.[10–14] Functional mutations disrupt specific interactions between the ligand and SH2 domain affecting specificity and activity, but have no impact on protein fold, whereas structural mutations damage the native structure and may lower or prevent expression of the protein. The production and stability of mRNA is essential as well. Analyses of the missense mutations among SH2 domains dictate the structural basis for the diseases as well as grant insight into the structure-function relationships of the SH2 domains.

Human disease-causing mutations typically affect amino acid positions conserved among protein families.[15] The positions have been conserved in evolution for protein function, stability and folding,[16,17] or for preventing aggregation.[18] The physical and chemical properties between the substituted residues have also been shown to be more diverse among the disease-causing mutations than for harmless substitutions, and disease-causing mutations often affect intrinsic structural features of proteins.[19–24] Furthermore, the severity of the substitution correlates with the likelihood of observing patients clinically.[15,25]

A generic registry of disease-causing mutations affecting SH2 domains was developed, and the mutations were collated also into corresponding locus-specific databases. The SH2base provides tools linking results from the mutational analyses to the locus-specific mutation data-

Table II
The Number of Reported Pathogenic Mutations in SH2 Domains[a]

	BTK	SH2D1A	SHP-2	RasGAP	P85α	ZAP-70	STAT1	STAT5B	Total
Missense	32/70	25/27	33/214	3/3	1/1	1/1	1/1	1/1	97/318
Nonsense	11/20	5/27							16/47
Insertion	4/4	2/2							6/6
Deletion	23/25	33/41					1/1		57/67
Splice site	19/23	11/11							30/34
Total	89/142	76/108	33/214	3/3	1/1	1/1	2/2	1/1	195/472

[a] The numbers are for unique mutations/unrelated families.

bases. The SH2base is freely available at http://bioinf.uta.fi/SH2base.

Several bioinformatic methods were applied to study the effects and consequences of each of the SH2 domain disease-related mutation. To analyze the putative structural effects of the pathogenic mutations, each missense mutation was modeled onto the corresponding structure and evaluated based on bioinformatic analysis. With the exception of STAT5B, the three dimensional structures of the SH2 domains studies here have been solved experimentally.[26–33] Furthermore, the inactive conformation of the SHP-2 protein was revealed by X-ray crystallography.[34] The majority of the mutations in all SH2 domains are located at positions involved in ligand binding and specificity, or in the case of SHP-2, at the interdomain interface. The disease-causing mutations were frequently found to affect conserved and covarying positions. We were able to elucidate the disease mechanism for each SH2 domain mutation.

MATERIALS AND METHODS

SH2base, a comprehensive collection of mutations affecting SH2 domains

The SH2base was built to provide a platform for the analyses of several different locus-specific mutation databases containing SH2 domain lesions. The SH2base web site includes results based on the structural and sequence analyses of mutations on corresponding protein structure. The database is freely available at http://bioinf.uta.fi/SH2base.

The patient data of each affected gene is stored in specialized locus-specific mutation databases. The disease-causing mutations identified in *BTK*, *SH2D1A*, *ZAP70*, *STAT1*, and *STAT5B* genes have been previously collated into BTKbase (http://bioinf.uta.fi/BTKbase/), SH2D1 Abase (http://bioinf.uta.fi/SH2D1Abase/), ZAP70base (http://bioinf.uta.fi//ZAP70base/), STAT1base (http://bioinf.uta.i/STAT1base/), and STAT5Bbase (http://bioinf.uta.fi/STAT5B base/), respectively.[35–37] The analyzed pathogenic mutations affecting *p85*α, *PTPN11*, and *RASA1* genes were collected from the literature into registries (http://bioinf.uta.fi/PIK3R1; http://bioinf.uta.fi/PTPN11base; http://bioinf.uta.fi/ RASA1base) according to the guidelines adopted in BTKbase.[36,38] The locus-specific SH2 databases are patient-based registries. Each entry contains four main items: identification of the patient and mutation, reference either to published article(s) or a submitting physician, mutation information, and data related to disease. The databases have a systematic design, which allows the use of the MUTbase program package[39] to generate new information and to distribute the data on the Internet. A submission program has been developed to submit mutation and patient information into databases.

Sequence alignments

For the sequence analysis, a ready-made sequence alignment of 1669 SH2 domain sequences was extracted from the Pfam database.[40] Structural alignments for the SH2 domains discussed here were made by the 3DCoffee server.[41]

Secondary structure assignment

The secondary structural boundaries were determined for each protein structure by the STRIDE algorithm.[42]

Determination of sequence conservation

For the analysis of sequence conservation, the entropy and information at each position or mutual information between any positions in the multiple sequence alignment was calculated as described in Ref. 43. For information calculations, amino acids were categorized into six physicochemically related groups: hydrophobic (V, I, L, F, M, W, Y, C), negatively charged (D, E), positively charged (R, K, H), conformational (G, P), polar (N, Q, S), and other (A, T). The threshold *P*-value for mutual information calculations was 0.01. The position-specific profiles were analyzed with MultiDisp (http://bioinf.uta.fi/cgi-bin/MultiDisp.cgi) (Riikonen *et al.*, in preparation) and ProCon.[43] The frequencies of residues in each alignment position were calculated, and when the frequencies of other residues than the consensus residue were zero or close to zero, the position was considered to be conserved. The degree of conservation, as presented in Figure 2(A), was determined by the ConSeq server, employing the empirical Bayesian statistics method for determining the conservation scores.[44] The analyses contained 1669 SH2 domain sequences extracted from the Pfam database.[40]

Analysis of structural effects

The consequences of missense mutations on protein structures were predicted based on structural analyses. The corresponding crystal structures were used in the analyses: SH2D1A (PDB code 1D4W),[30] ZAP-70 N-terminal SH2 domain (1M61),[27] SHP-2 (2SHP),[34] SH2 domain of p85α (2IUG),[45] STAT1 (1YVL),[46] and the solution structures of the BTK SH2 domain (2GE9),[33] and RasGAP (2GSB).

Mutations either affecting proline or introducing a proline in the ordered secondary structures were predicted to cause backbone strain. Overpacking was measured by rotating each of the mutated residues over full range of side chain χ angles. Only the substituted side chain was allowed to move during the analyses. The rotatable side chain was created with PREKIN 5.93 and an automated sampling of torsional angles was done with the Autobondrot procedure under PROBE 2.80.[47] The acceptable conformations for a mutated side chain have a total score of above −1.0 allowing for small local perturbations in the structure.[48] Mutations with possible rotamer conformations were modeled on corresponding wild type structure. The best rotamer was used to build the mutated amino acid. The hydrogens were added by using Insight II (Accelrys) or Reduce.[49] The models were then analyzed for positive van der Waals surfaces and electrostatic effects by using PROBE and MAGE programs,[47] or the PyMOL program.[50] The changes in hydrogen bonds, accessible surface area, and salt bridges affecting contacts between residues were calculated with program WhatIf,[51] RankViaContact,[52] and CSU services.[53] The ranked contact energies were used to estimate the importance of the residue to the stability of the protein structure. The electrostatic surface potentials were calculated and visualized with PyMOL.

RESULTS AND DISCUSSION

We have collated all reported pathogenic mutations affecting SH2 domains into a database (http://bioinf.uta.fi/SH2base), and analyzed the consequences of missense mutations on each affected protein by bioinformatics methods.

The SH2 domains form a distinct well-characterized protein domain family affected by a wide range of pathogenic mutations. SH2 domain is a relatively small protein interaction domain that folds into a distinct antiparallel β-sheet structure sandwiched between two α-helices. The structure has an elongated binding site for phosphotyrosine-containing peptides or proteins perpendicular to the β-sheet. The SH2 domains are usually found in multidomain proteins, together with kinase or phosphatase domains and additional domains involved in mediating the assembly of specific protein complexes (see Fig. 1). SH2D1A is a small protein consisting solely of the SH2 domain with a C-terminal extension, whereas all the other SH2-domain containing proteins discussed here consist of several domains. There are two SH2 domains in ZAP-70, SHP-2, RasGAP, and p85α, but only in SHP-2 there are mutations in both the N- and C-terminal SH2 domains.

The SH2 domains in BTK, SH2D1A, SHP-2, p85α, RasGAP, ZAP-70, STAT1, and STAT5B are known to be affected in different types of immunodeficiencies and cancer, insulin deficiency, or growth hormone insensitivity (Table I). Currently, 195 unique molecular events in 472 unrelated patients have been reported (Table II). A generic registry of disease-causing mutations affecting SH2 domains was developed. In addition, all pathogenic mutations were collated into corresponding locus-specific mutation databases. The SH2base provides tools combining mutation analyses to the particular locus-specific mutation database describing the defective gene and patient data.

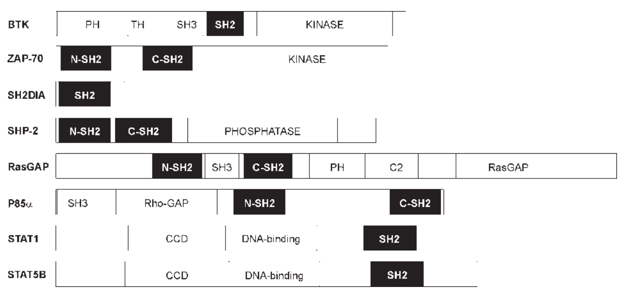


Figure 1

The domain organization of the analyzed proteins. The SH2 domains in each protein are highlighted black. TH, Tec homology domain; CCD, coiled coil domain.

General overview of SH2 domain mutations

We have analyzed missense mutations in all eight affected SH2 domains by using bioinformatic methods on sequence and structural level. We have applied this approach into the analysis of numerous diseases, for example, Refs. 37,54–58. The sequence conservation of each position was determined at three levels: identity, conservation of amino acid physicochemical properties, and covariant conservation. Positions involved in ligand binding were elucidated from literature. Mutations causing steric clashes and over-packing were analyzed by rotating the introduced side chain χ-angles. In case the introduced residue fitted on the structure, the model was analyzed for loss of any favorable interactions to interpret the putative consequences of a particular mutation.

The vast majority of the SH2 domain affecting mutations is found in the *BTK* and *SH2D1A* genes, whereas in p85α, RasGAP, ZAP-70, STAT1, and STAT5B there are only few known mutations in each gene (Table II). In SHP-2, all the known disease-causing mutations are missense mutations, which is in agreement with the gain of function role of these mutations. The 29 unique intron mutations affecting SH2 domains in 36 unrelated patients cause aberrant splicing and lead to altered protein products or induce changes in transcriptional activity. Short deletions are more frequent than insertions as has been shown for many diseases.[59] The 16 different nonsense mutations affect functional and structural positions in the SH2 domain structures. Of the 100 different missense mutations, 46% are located on β-strands followed by loop structures (27%), and α-helices (26%). The single STAT5B mutation was excluded from this analysis due to uncertain secondary structure alignment.

The majority of the disease-causing missense mutations in SH2 domains affect functionally important amino acids (71%) that are involved either in the ligand binding or interactions between the N-terminal SH2 domain of SHP-2 and the phosphatase domain. Based on our analyses of the SH2 domain structures, 29% of missense mutations were found to affect protein stability, with backbone strain (12% of all the mutations), and over-packing (8% of all the mutations) being the most common defects. Most of the pathogenic missense mutations affecting corresponding positions in several SH2 domains are located on the binding surface. These missense mutations affect highly conserved residues involved in phosphotyrosine binding as well as unconserved positions related to SH2 domain specificity [Fig. 2(B,C)]. Furthermore, the probability of observing a covariant pair when either of the residues is mutated is extremely low. Many mutations also alter the electrostatic surface potential of the protein, especially at the phosphotyrosine binding pocket [Fig. 2(D,E)]. Electrostatic surface potentials are crucial in ligand binding, and the surface charge–charge relationships have a role in maintaining the stability of the protein as well.[60]

The highest numbers of mutations appear in BTK, SH2D1A, and the N-terminal SH2 domain of SHP-2. The types of mutations differ among the three proteins. In BTK, most missense mutations are located at the ligand-binding surfaces, and can thus be regarded as functional, having no impact on protein fold. The SHP-2 mutations are functional as well, most of them being located at the NSH2-PTP-interacting surface, thereby regulating the protein activity. In contrast, the SH2D1A mutations are mostly structural, affecting buried residues in the protein core and leading to over-packing.

All eight analyzed SH2 domain structures shared the typical SH2 domain fold. For the general nomenclature of the mutated sites the system introduced by Eck *et al*.[61] was used. This system is based on the secondary structures and regions connecting them (see Fig. 3). The βA, βB, βC, and βD strands form an antiparallel β-sheet with two α-helices on both sides of the β-sheet.[62] The phosphotyrosine-binding pocket is formed by amino acids located in the αA helix, βB and βC strands, and in the BC-loop [Fig. 2(A)]. Residues from αB-helix and the EF and BG-loops are involved in binding of amino acids C-terminal to the phosphotyrosine in the ligand. The βD', βE, and βF strands form an additional small β-sheet that closes off one part of this side [Fig. 2(A)].

Mutations in different diseases

Missense mutations affecting SHP-2 are mainly located on the surface between the N-terminal SH2 domain and the phosphatase domain. The mutations favor the active enzyme conformation and the wide spectrum of phenotypes caused by *PTPN11* mutations results from a constantly active phosphatase in cells.[63] In agreement, mutations leading to the depletion of enzyme activity, such as nonsense mutations or large deletions, have not been identified from the NS patients. Four missense mutations affect positions located in the phosphopeptide-binding clefts (T42A, L43F, and T52S in the N-terminal SH2 domain, and E139D in the C-terminal SH2 domain), and are thus predicted to perturb phophopeptide binding specificity or/and affinity (Table III). Two of these mutations have been shown to increase SHP-2 activity after stimulation compared with the wild type protein.[64] In contrast to *PTPN11*, defects affecting *BTK* and *SH2D1A* genes range from gross deletions of the whole gene to single point mutations.

The effects of several XLP causing missense mutations on protein stability and ligand binding have been studied *in vitro* and *in vivo*.[12,28,65–67] As expected, mutations involved directly in ligand binding appear to be stable *in vivo* abolishing the interaction between the SH2D1A and the cytoplasmic tail of SLAM family members. Mutations outside ligand-binding areas were shown to share a significantly reduced half-life indicating more rapid proteolysis of the mutated proteins inside the cells. The SH2 domain of BTK has also been found to tolerate XLA causing mutations poorly *in vitro*, although the mutated full-length protein appears to be more stable *in vivo*.[11]

The missense mutations in BTK mainly affect surface-exposed residues involved in phophopeptide binding and specificity. Interestingly, almost all the mutant side chains could be adopted into the structure without major structural rearrangements. Substitutions introducing a proline are quite common, and these are predicted to cause alterations in the structure, since proline is a known secondary structure breaker. Mutations to/from proline are usually deleterious.[68] One of the proline substitutions is at the BG-loop, which is involved in ligand binding in all SH2 domains. The RasGAP mutations are all located at the specificity-determining βD-strand, and are expected to have a functional role.

The single SCID-causing ZAP-70 missense mutation P80Q has been demonstrated to have a structural, rather than functional role. The mutant protein is unstable *in vivo*, leading to decreased half-life and rapid degradation of the protein.[69]

The STAT1 and STAT5B mutations both introduce prolines in β-strands. In STAT1, the L600P mutation is positioned close to the phosphopeptide-binding R602 (RβB4), and the inserted proline could alter the orientation of the strand as a result of the introduced backbone strain, thereby changing the position of the critical arginine. The leucine and the affected βB-strand amino acids are conserved in all STATs. On the other hand, the mutant protein could not be detected in patient cells,[70] which suggests that the protein is not stable and it is quickly degraded after synthesis. Thus, the introduced proline would not only prevent the function of the protein by disrupting the local structure, but also cause more extensive structural damage to the protein. The STAT5B mutation, A630P, is located at βC5, by homology. This position probably has a role in maintaining the

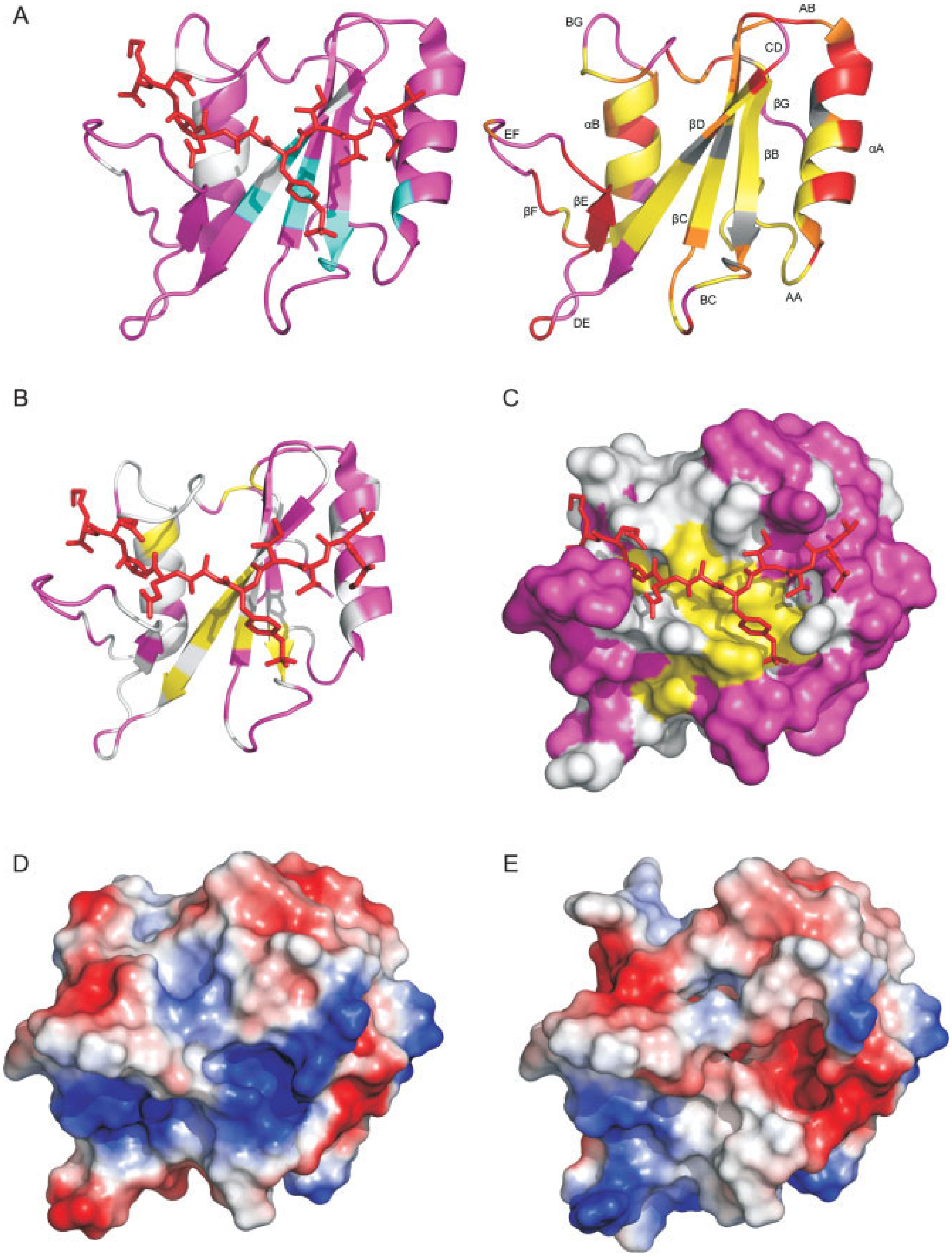
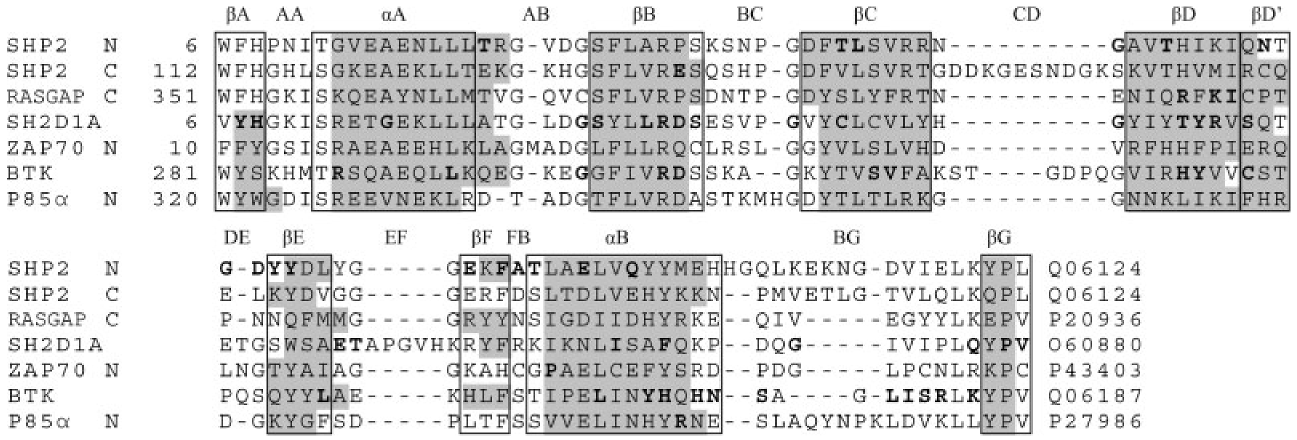


Figure 2

(*A*) *Stereo view of SH2D1A. On the left, the residues involved in phosphotyrosine binding in any of the SH2 domains discussed are colored cyan. Residues involved in specificity are colored white. On the right, the color coding refers to sequence conservation in SH2 domains. The most conserved positions are colored gray followed by yellow, orange, magenta, and the most variable regions are colored red. The secondary structures are indicated.* (***B***) *The SH2 domain of SH2D1A (magenta) in complex with a SLAM receptor derived phosphopeptide ligand (red). All known pathogenic missense mutations reported from the SH2 domains are shown. Mutations affecting only one SH2 domain are shown in white, whereas mutations found in more than one SH2 domain are in yellow.* (***C***) *Surface representation of the SH2 domain of SH2D1A showing the two binding pockets. Mutations are colored as in Figure 2(B).* (***D***) *Electrostatic surface potential of wild type SH2D1A. The blue color indicates positive potential, and the red color indicates negative potential.* (***E***) *Electrostatic surface potential of mutated SH2D1A. All the missense mutations affecting any of the analyzed SH2 domains are included. The mutations have a net effect of changing the surface potential more negative, especially at the phosphotyrosine binding pocket and its surroundings.*



```
                  βA  AA        αA          AB       βB        BC        βC              CD            βD     βD'
SHP2    N     6   WFHPNITGVEAENLLLTRG-VDGSFLARPSKSNP-GDFTLSVRRN----------GAVTHIKIQNT
SHP2    C   112   WFHGHLSGKEAEKLLTEKG-KHGSFLVRESQSHP-GDFVLSVRTGDDKGESNDGKSKVTHVMIRCQ
RASGAP  C   351   WFHGKISKQEAYNLLMTVG-QVCSFLVRPSDNTP-GDYSLYFRTN----------ENIQRFKICPT
SH2D1A        6   VYHGKISRETGEKLLLATG-LDGSYLLRDSESVP-GVYCLCVLYH----------GYIYTYRVSQT
ZAP70   N    10   FFYGSISRAEAEEHLKLAGMADGLFLLRQCLRSL-GGYVLSLVHD----------VRFHHFPIERQ
BTK         281   WYSKHMTRSQAEQLLKQEG-KEGGFIVRDSSKA--GKYTVSVFAKST----GDPQGVIRHYVVCST
P85α    N   320   WYWGDISREEVNEKLRD-T-ADGTFLVRDASTKMHGDYTLTLRKG----------GNNKLIKIFHR

                  DE  βE       EF      βF FB     αB               BG         βG
SHP2    N         G-DYYDLYG-----GEKFATLAELVQYYMEHHGQLKEKNG-DVIELKYPL  Q06124
SHP2    C         E-LKYDVGG-----GERFDSLTDLVEHYKKN--PMVETLG-TVLQLKQPL  Q06124
RASGAP  C         P-NNQFMMG-----GRYYNSIGDIIDHYRKE--QIV-----EGYYLKEPV  P20936
SH2D1A            ETGSWSAETAPGVHKRYFRKIKNLISAFQKP--DQG-----IVIPLQYPV  O60880
ZAP70   N         LNGTYAIAG-----GKAHCGPAELCEFYSRD--PDG-----LPCNLRKPC  P43403
BTK               PQSQYYLAE-----KHLFSTIPELINYHQHN--SA----G-LISRLKYPV  Q06187
P85α    N         D-GKYGFSD-----PLTFSSVVELINHYRNE--SLAQYNPKLDVKLLYPV  P27986
```

Figure 3

Structural sequence alignment of the defective SH2 domains, excluding STAT1 and STAT5B. The boxes indicate common secondary structural elements used for defining the nomenclature in the text and in the Figures 2 and 4. The gray highlights indicate the secondary structural boundaries for each protein, calculated from atomic coordinates of each protein by the program STRIDE. The positions of disease-causing mutations are shown in bold. SwissProt codes are given after corresponding sequence.

phosphotyrosine binding pocket structure (see below). The mutation has been shown to cause loss of thermodynamic stability as well as aberrant folding and aggregation of the protein.[71] Interestingly, in XLA the βC5 position is mutated to proline as well.

Amino acid conservation

Amino acid conservation was determined with the program ProCon on three levels, namely invariant positions (only RβB5), physicochemically conserved sites, and the network formed by covariant residues (see Fig. 4). The degree of sequence conservation, as determined by the program ConSeq, is illustrated in Figure 2(A). The most conserved sites are generally the ligand-binding residues and the residues surrounding them, whereas the most variable regions are situated at the outer edges of the helices, and loop regions. In BTK, 48% of the mutations affect conserved positions, most of which are functional positions involved in phosphopeptide binding. The number is about the same in SH2D1A, where 46% of the mutations affect conserved positions. The conserved positions in SH2D1A are mainly structural, and many mutations affect covariant positions. Only four of the SHP-2 mutation positions are conserved, two are involved in phosphopeptide binding and the other two are positioned at the N-SH2-PTP-interaction surface.

The majority of SHP-2 missense mutations are located at the interdomain interface, thereby destabilizing the inactive structure. This mechanism of regulation of activity, exhibited by SHP-1 and SHP-2, is unique among the SH2 domains,[72] and therefore positions involved in the interdomain binding are generally not conserved among the SH2 family. In the αB2 position, which is mutated in ZAP-70 (P80Q), hydrophobicity is a conserved property, and no charged residues appear among the SH2 domains. Of the three mutated positions in RasGAP, two are conserved (βD4 and βD7). The STAT1 L600 is conserved among the STAT proteins (see Fig. 5).

There are 19 covariantly conserved positions in the SH2 family (see Fig. 4), 11 of which are affected by the disease-causing mutations. The covarying positions βB6, βC3, βD4 are all mutated in SH2D1A, and there are mutations in these positions in at least one other SH2 domain, as well. Almost every position in the βC strand is covariantly conserved, excluding βC1. This reflects the important role of the βC strand in forming the phosphotyrosine binding pocket together with residues from the βB and βD strand [Fig. 2(A,B)]. For example, the positions βB4, βC2, βC4, and βD7 form a separate covarying group (see Fig. 4), and are involved in a network of hydrophobic contacts in the protein core right below the phosphotyrosine-binding pocket.

Mutations affecting phosphotyrosine binding

The residues interacting with the phosphotyrosine are generally conserved and form a positively charged binding pocket on the SH2 domain surface.[62] Thirty-four different missense mutations in 23 positions presumably affect phosphotyrosine binding. Five of these positions showed statistically significant covariance ($P < 0.01$) with at least one other position (see Fig. 4).

Table III
Summary of the Mutations and their Effects

Protein	Mutation	Position	Structural[a]	Functional[b]	Phenotype
BTK	R288Q	αA2		X	Classical XLA
	R288W	αA2		X	Mild/classical XLA
	L295P	αA9	X		Classical XLA
	G302E	AB6		X	Classical XLA
	G302R	AB6		X	Classical XLA
	R307G	βB5		X	Classical XLA
	R307K	βB5		X	Classical XLA
	R307T	βB5		X	Classical XLA
	D308E	βB6		X	XLA
	S318F	βC5		X	Classical XLA
	S318P	βC5		X	Classical XLA
	V319A	βC6	X		Classical XLA
	Y334S	βD5		X	Classical XLA
	C337G	βD′1	X		Classical XLA
	L346R	βE4		X	Classical XLA
	L346P	βE4	X		Classical XLA
	L358F	αB5		X	Classical XLA
	Y361C	αB8		X	XLA
	Y361D	αB8		X	Classical XLA
	H362Q	αB9		X	Classical XLA
	H362R	αB9		X	XLA
	H364D	αB11		X	Classical XLA
	H364P	αB11	X		XLA
	N365Y	αB12	X		Classical XLA
	S366F	BG1	X		XLA
	L369F	BG4		X	Classical XLA
	I370M	BG5		X	XLA
	S371P	BG6	X		Classical XLA
	R372G	BG7	X		XLA
	K374N	BG9	X		Classical XLA
SH2D1A	Y7C	βA2	X		XLP
	H8D	βA3	X		XLP
	H8P	βA3	X		XLP
	G16D	αA5		X	XLP
	G27S	AB6	X		XLP
	S28R	βB1	X		XLP
	L31P	βB4	X		XLP
	R32T	βB5		X	XLP
	D33Y	βB6		X	XLP
	S34G	βB7		X	XLP
	S34R	βB7		X	XLP
	G39V	BC5	X		XLP
	C42W	βC3		X	XLP
	G49V	CD2	X		XLP
	T53R	βD4		X	XLP
	Y54C	βD5		X	XLP
	R55L	βD6		X	XLP
	S57P	βD′1	X		XLP
	E67D	EF1		X	XLP
	T68I	EF2		X	XLP
	I84T	αB6	X		XLP
	F87S	αB9	X		XLP
	G93D	BG3		X	XLP
	Q99P	BG9	X		XLP
	P101L	βG2	X		XLP
	V102G	βG3	X		XLP
P85α	R409Q	αB10	X		Severe insulin deficiency
ZAP-70	P80Q	αB2	X		SCID
SHP-2	T42A	βC3		X	NS
	L43F	βC4		X	NS
	T52S	βD3		X	JMML
	N58D	βD′2		X	NS
	N58Y	βD′2		X	ALL

Table III
(Continued)

Protein	Mutation	Position	Structural[a]	Functional[b]	Phenotype
	N58H	βD′2		X	NS
	N58K	βD′2		X	NS
	G60R	DE1		X	JMML
	G60V	DE1		X	MDS/JMML
	G60A	DE1		X	NS
	D61G	DE2		X	NS/JMML/MPD
	D61N	DE2		X	NS/JMML
	D61Y	DE2		X	ALL/JMML
	D61V	DE2		X	ALL/JMML/MDS
	Y62D	βE1		X	NS/JMML
	Y62N	βE1		X	NS
	Y63C	βE2		X	NS
	E69Q	βF1		X	NS
	E69K	βF1		X	ALL/JMML/MDS
	F71L	βF3		X	MDS/NS
	F71K	βF3		X	AML
	A72G	FB1		X	NS
	A72S	FB1		X	NS
	A72T	FB1		X	ALL/JMML
	A72V	FB1		X	ALL/JMML/AML
	A72D	FB1		X	ALL
	T73I	αB1		X	NS/JMML/MPD
	E76Q	αB4		X	ALL/JMML
	E76K	αB4		X	ALL/AML/JMML
	E76V	αB4		X	JMML
	E76G	αB4		X	ALL/JMML
	E76A	αB4		X	JMML/MDS
	E76D	αB4		X	NS
	E76M	αB4		X	NS
	Q79P	αB7		X	NS
	Q79R	αB7		X	NS
	E139D	βB6		X	NS/JMML
RasGAP	R398L	βD4		X	BCC
	K400G	βD6		X	BCC
	I401V	βD7	X		BCC
STAT1	L600P	βB2	X		STAT1 deficiency
STAT5b	A630P	βC5	X		Growth hormone insensitivity

[a] Mutations that do not fit into the structure or cause loss of important contacts between residues are designated as structural.
[b] Mutations that affect ligand-binding residues or SHP-2 residues involved in N-SH2-PTP interaction surface are considered functional.

The most frequently mutated positions among the XLA and XLP patients are located in the phosphotyrosine binding-pocket. The second arginine at the αA helix is conserved in almost all SH2 domains and it is involved in coordinating the phosphotyrosine. The R288Q and R288W (αA2) mutations in BTK disturb phosphopeptide binding,[33] although they both can be fitted into the binding pocket. Both mutations at αA2 have also been shown to cause severe loss of binding affinity in biochemical studies.[73] In the SH2D1A structure, the phosphate group from the phosphotyrosine is rotated and the RαA2 interactions are replaced by R55 (βD6). XLP patients have two different mutations at this position; the codon is either changed to leucine or termination codon. The leucine mp-rotamer conformation can be fitted into the binding pocket, but the coordinating interactions to the phosphate group are lost. The XLA and XLP mutations in the αA2 and βD6 positions contain CpG sites at the DNA level. Methylated cytosines appear with high frequency in CpG dinucleotides, and spontaneous deamination of methylcytosine to thymine underlies in many primary immunodeficiencies[74] or in different types of cancer.[75] The majority of CpG mutations affect arginine residues. In many cases, the substituting residues cannot replace the long and charged side chain without affecting protein function.[76–78]

The arginine at the fifth position of the βB strand (RβB5) is the only invariant residue among the SH2 domains. This arginine is situated in the bottom of the phosphotyrosine-binding pocket and is responsible for binding and identification of the phosphorylated tyrosine of the ligand.[62] The isothermal titration calorimetry binding studies with high affinity show that the phosphotyrosyl group contributes about 60% of the free energy of the interaction.[79,80] In five unrelated patients

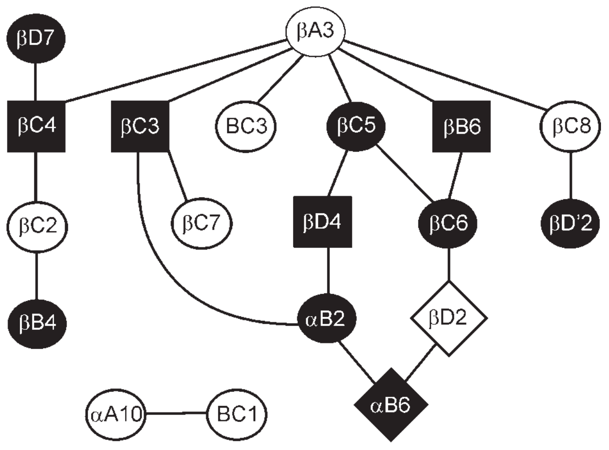


Figure 4

The positions among the SH2 domains showing statistically significant covariance. Positions involved in binding of the phosphotyrosine in any of the analyzed SH2 domains are inside squares, whereas positions related to specificity are inside diamonds. Positions carrying disease-causing mutations are indicated by a black background.

with XLA, the RβB5 position has been mutated to glycine, lysine, or threonine, whereas the XLP patient has threonine. In both domains, the substitutions cause the disease by removing the crucial interaction required for the recognition and binding of the phosphotyrosine. In agreement with the predictions, the XLA causing mutation to glycine was shown to abolish binding of phosphotyrosine *in vitro*, although the mutated and wild type proteins had identical CD spectra.[11] The binding affinities of the mutants introducing glycine or threonine at RβB5 were also shown to be greatly reduced.[73] The patient monocytes expressing the R307G mutant also had normal amount of BTK protein.[81] Mutation of the arginine to lysine or alanine has been shown to abrogate the regulatory role of BTK in sustained intracellular Ca^{2+} signaling and Fas-mediated apoptosis in B cells.[82]

In addition to defects at the βB5 position, other mutations affecting residues involved in the formation of the phosphotyrosine binding pocket cause functional defects in the protein, without necessarily causing large perturbations into the structure. As an example, the conserved βD4 position is typically occupied by an amino acid with a large hydrophobic moiety that can interact with the phenol ring of phosphotyrosine.[62] The position shows significant covariation with βC5 (see Fig. 4), a position that most probably has a role in stabilizing the structure of the binding pocket, since the two residues are in extensive contact. In BTK and RasGAP (C-SH2), the βD4 position is occupied by a histidine and arginine, respectively, whereas SH2D1A has threonine. In the RasGAP SH2 domain, the BCC-causing R398L mutation at this position can be fitted into the structure, but numerous hydrogen bonds formed by the arginine are lost as a consequence of the substitution, including the hydrogen bond with Y389 in the covarying βC5 position. The XLP causing substitution H333L at the corresponding position can be accommodated into the structure as well, but crucial hydrogen bonds with residues 288 and 292 are lost, which is predicted to lead to alterations in the orientation of the pY-binding R288. The SH2D1A βD4 position, on the contrary, is solvent accessible. The T53 at this position interacts with the pY +2 residue from the ligand.[28,30] Biochemical analyses have shown that the mutated protein is stable *in vivo* and the substitution only abolishes binding to the unphosphorylated SLAM receptor.[12] Based on the SH2D1A structure with target peptide (1D4W), the side chain of isoleucine collides with the two residues preceding phosphotyrosine in all rotamer conformations indicating a dynamic process at the binding surface that our method cannot predict.

Mutations affecting SH2 domain specificity

The residues involved in binding of the third residue following phosphotyrosine (pY +3) are located in the

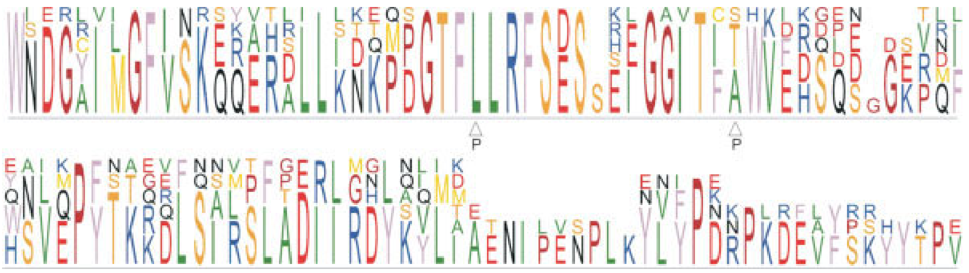


Figure 5

A MultiDisp visualization of the sequence alignment for the SH2 domains of the STAT family of proteins (human STAT1, STAT2, STAT3, STAT4, STAT5A, STAT5B, and STAT6). The height of the characters indicates the frequency of the amino acids in the alignment positions, and the color of the objects represents the chemical nature of the amino acids (physicochemically related amino acids have the same color). Arrowheads below the alignment, together with the mutant forms, indicate the positions of STAT SH2 domain mutations.

αB helix and in the EF and BG loops. These residues are highly variable and respond to individual SH2 domain specificity [Fig. 2(A,B)]. In the SH2 domains of BTK, SH2D1A, p85α, ZAP-70, and RasGAP, the ligand-binding residues come close together forming another binding pocket on the SH2 domain surface.[14,27,30,32] In SHP-2, the residues move away from each other opening up a binding groove on the SH2 domain surface and let the interactions between the ligand and SH2 domains extend beyond the pY +3 position.[29] The first residue following the phosphotyrosine (pY +1) is bound on the surface between the two binding pockets. In STAT1, the phosphopeptide-SH2 interaction is primarily attributed to residues pY, pY +1, and pY +4 of the peptide.[46]

The majority of the peptide-binding motifs for individual SH2 domains have been identified by using *in vitro* oriented phosphopeptide library assays.[83] Based on these results, together with structural analyses of different ligand-binding models, the βD5 position has been shown to play a crucial role in the determination of specificity. In BTK and SH2D1A SH2 domains, the tyrosine forms part of the pY +3-binding pocket and interact with the pY +1 residue from the ligand as well. The XLA (Y334S) and XLP (Y54C) mutations at this position affect specific interactions between the SH2 domain and the ligand. The BTK mutation Y334S was shown to reduce binding of the wild type preferred ligands, and the mutated protein preferred hydrophobic residues at pY +1 position.[73] Furthermore, transient expression of the mutated full-length protein in COS-7 cells showed wild-type stability and was not prone to aggregation.[11] The XLP causing mutation Y54C reduces the thermodynamic stability and affinity to the SLAM receptor derived peptide.[65] The surrounding residues βD4 and βD6, which are involved in forming the binding pocket, are affected by disease-causing missense mutations as well. The βD4 mutations in BTK, SH2D1A, and RasGAP all cause a change of charge in the binding pocket, which likely affects ligand binding as well. There are mutations affecting the βD6 position in RasGAP and SH2D1A, and these substitutions affect the charge of the binding pocket as well.

Several residues forming the hydrophobic pocket for the ligand pY +3 residue have been mutated causing XLA and XLP (Y334S, Y361C, Y361D, L369F, and I370M in XLA, and E67D, T68I, and G93D in XLP). The majority of these defective residues in BTK are located in the N-terminus of the αB helix or in the BG-loop. In contrast, the XLP causing mutations are located in the EF and BG loops. There are two XLP mutations in the αB helix: I84T (αB6) and F87S (αB9). The threonine 84 can be accommodated into the structure, the best rotamer having almost identical positions for Cβ and Cγ as the wild type isoleucine. The mutation has been shown to cause reduced half-life of the protein, and the ability of the mutated protein to induce downstream signaling through SLAM is approximately fourfold less than that of the wild type protein in spite that the mutated protein was able to bind SLAM.[66] The increase in polarity of the buried side chain probably causes loss of structural stability in the protein. In the best rotamer conformation, the Cβ and the hydroxyl group of the serine 87 were in identical conformation as the Cβ and Cγ of the wild type phenylalanine, but the hydroxyl group of the serine is oriented into the hydrophobic binding pocket. The XLP causing mutation was shown to decrease binding affinity and thermal stability *in vitro*.[65] The H362Q and H362R (αB9) mutations in BTK are likely to disrupt the binding pocket conformation. In agreement with the predictions, the H362Q mutation has been shown to abolish phosphotyrosine binding *in vitro* and the mutated protein has an nonnative-like CD spectrum indicating changes in the secondary structures in relation to each other.[11]

Genotype-phenotype correlations

In XLA, there most likely are some genotype–phenotype (GP) correlations, because the severity of the disease varies among patients.[84] In the severe (classical) form of XLA, the susceptibility to severe infections is greater and the levels of B-lymphocytes and/or immunoglobulin are lower than in the mild form. The age at diagnosis also corresponds to the severity of the disease. However, the GP correlations are difficult to elucidate, since the same molecular event may result in different forms of the disease in different families, or the phenotype may vary even within certain kindreds.[84]

GP correlations have also been found in *PTPN11* mutations with patients with NS, JMML, and NS/MPD.[85] The somatically acquired JMML-causing mutations have been suggested to have a strong gain-of-function effect that might affect embryonic or fetal development if they were transmitted in the germ line, which is why these mutations are not observed in NS patients. The hereditary NS-causing mutations are supposed to have a weaker effect, and the effects of NS/MPD-causing mutations would have intermediate effects on the protein function.[85] For example, the NS causing mutations N58D, G60A, D61N, E69Q, A72G, and E76D could be considered to have a mild effect on the protein, since the substituting amino acid resembles the wild type amino acid biochemically. Yet there are cases where a more pronounced change in amino acid properties causes NS, and some substitutions predicted to have a mild effect lead to JMML or AML (e.g., E76Q). However, the dataset of 35 individual substitutions is not sufficient for a systematic statistical study of the GP correlations. Similarly as in XLA, the GP correlations in *PTPN11* mutations are not clear because some mutations cause different phenotypes in different families. For example, the mutations D61G and T73I cause NS, NS/MPD, and JMML in different kindreds.

CONCLUSIONS

We have shown that many pathogenic mutations affecting SH2 domains are located either in strictly conserved positions or influence covariant pairs. In general, a disease-causing mutation at a covariant pair decreases the probability of finding the mutated residue and its complementing residue to zero. The pathogenic mutations were found to frequently affect covarying pairs among the SH2 domains.

The majority of the substitutions in the SH2 domains affect residues involved in ligand binding and specificity. In the SHP-2 SH2 domain, 19 of the 24 mutations are positioned at the N-SH2-PTP-interacting surface, thereby involved in the regulation of protein activity. Four of the mutations affect phosphotyrosine binding. Few XLA causing mutations were predicted to lead into dramatic changes in the protein structure. Twenty of the mutations in BTK affect residues involved in ligand binding, revealed by biochemical studies,[11,73] and the NMR structure.[33] The mutations in RasGAP also cluster in and around the ligand-binding pocket. There are more structural mutations in SH2D1A, leading to overpacking and loss of favorable interactions maintaining the proper structure. In the proteins with only one known disease-causing mutation, the effects are predicted to be structural, because they all, excluding p85α, affect or introduce prolines. The structural effects of these mutations have also been demonstrated in biochemical studies.[69,70,86] Clearly, more biochemical analyses are required to understand the biophysical properties of each defective SH2 domain to test our predictive findings, and to fully understand the molecular mechanisms behind the diseases.

REFERENCES

1. Schlessinger J, Lemmon MA. SH2 and PTB domains in tyrosine kinase signaling. Sci STKE 2003;191:RE12.
2. Machida K, Mayer BJ. The SH2 domain: versatile signaling module and pharmaceutical target. Biochim Biophys Acta 2005;1747:1–25.
3. Nars M, Vihinen M. Coevolution of the domains of cytoplasmic tyrosine kinases. Mol Biol Evol 2001;18:312–321.
4. Chan AC, Iwashima M, Turck CW, Weiss A. ZAP-70: a 70 kd protein-tyrosine kinase that associates with the TCR ζ chain. Cell 1992;71:649–662.
5. Engel P, Eck MJ, Terhorst C. The SAP and SLAM families in immune responses and X-linked lymphoproliferative disease. Nat Rev Immunol 2003;3:813–821.
6. Friedman E. The role of ras GTPase activating protein in human tumorigenesis. Pathobiology 1995;63:348–350.
7. Kurosaki T. Regulation of B-cell signal transduction by adaptor proteins. Nat Rev Immunol 2002;2:354–363.
8. Neel BG, Gu H, Pao L. The 'Shp'ing news: SH2 domain-containing tyrosine phosphatases in cell signaling. Trends Biochem Sci 2003;28: 284–293.
9. Shepherd PR, Withers DJ, Siddle K Phosphoinositide 3-kinase: the key switch mechanism in insulin signalling. Biochem J 1998;333: 471–490.
10. Nera KP, Brockmann E, Vihinen M, Smith CIE, Mattsson PT. Rational design and purification of human Bruton's tyrosine kinase SH3-SH2 protein for structure-function studies. Protein Expr Purif 2000;20:365–371.
11. Mattsson PT, Lappalainen I, Backesjö C-M, Brockmann E, Laurén S, Vihinen M, Smith CIE. Six X-linked agammaglobulinemia-causing missense mutations in the Src homology 2 domain of Bruton's tyrosine kinase: phosphotyrosine-binding and circular dichroism analysis. J Immunol 2000;164:4170–4177.
12. Morra M, Simarro-Grande M, Martin M, Chen AS, Lanyi A, Silander O, Calpe S, Davis J, Pawson T, Eck MJ, Sumegi J, Engel P, Li SC, Terhorst C. Characterization of SH2D1A missense mutations identified in X-linked lymphoproliferative disease patients. J Biol Chem 2001;276:36809–36816.
13. Bocchinfuso G, Stella L, Martinelli S, Flex E, Carta C, Pantaleoni F, Pispisa B, Venanzi M, Tartaglia M, Palleschi A. Structural and functional effects of disease-causing amino acid substitutions affecting residues Ala72 and Glu76 of the protein tyrosine phosphatase SHP-2. Proteins 2006;66:963–974.
14. Vihinen M, Nilsson L, Smith CIE. Structural basis of SH2 domain mutations in X-linked agammaglobulinemia. Biochem Biophys Res Commun 1994;205:1270–1277.
15. Miller MP, Kumar S. Understanding human disease mutations through the use of interspecific genetic variation. Hum Mol Genet 2001;10:2319–2328.
16. Hamill SJ, Cota E, Chothia C, Clarke J. Conservation of folding and stability within a protein family: the tyrosine corner as an evolutionary cul-de-sac. J Mol Biol 2000;295:641–649.
17. Mirny LA, Shakhnovich EI. Universally conserved positions in protein folds: reading evolutionary signals about stability, folding kinetics and function. J Mol Biol 1999;291:177–196.
18. Steward A, Adhya S, Clarke J. Sequence conservation in Ig-like domains: the role of highly conserved proline residues in the fibronectin type III superfamily. J Mol Biol 2002;318:935–940.
19. Chasman D, Adams RM. Predicting the functional consequences of non-synonymous single nucleotide polymorphisms: structure-based assessment of amino acid variation. J Mol Biol 2001;307:683–706.
20. Ferrer-Costa C, Orozco M, de la Cruz X. Characterization of disease-associated single amino acid polymorphisms in terms of sequence and structure properties. J Mol Biol 2002;315:771–786.
21. Wang Z, Moult J. SNPs, protein structure, and disease. Hum Mutat 2001;17:263–270.
22. Sunyaev S, Ramensky V, Bork P. Towards a structural basis of human non-synonymous single nucleotide polymorphisms. Trends Genet 2000;16:198–200.
23. Khan S, Vihinen M. Spectrum of disease-causing mutations in protein secondary structures. BMC Struct Biol 2007;7:56.
24. Sunyaev S, Ramensky V, Koch I, Lathe W, III, Kondrashov AS, Bork P. Prediction of deleterious human alleles. Hum Mol Genet 2001;10:591–597.
25. Steward RE, MacArthur MW, Laskowski RA, Thornton JM. Molecular basis of inherited diseases: a structural perspective. Trends Genet 2003;19:505–513.
26. Eck MJ, Pluskey S, Trub T, Harrison SC, Shoelson SE. Spatial constraints on the recognition of phosphoproteins by the tandem SH2 domains of the phosphatase SH-PTP2. Nature 1996;379:277–280.
27. Folmer RH, Geschwindner S, Xue Y. Crystal structure and NMR studies of the apo SH2 domains of ZAP-70: two bikes rather than a tandem. Biochemistry 2002;41:14176–14184.
28. Hwang PM, Li C, Morra M, Lillywhite J, Muhandiram DR, Gertler F, Terhorst C, Kay LE, Pawson T, Forman-Kay JD, Li SC. A "three-pronged" binding mechanism for the SAP/SH2D1A SH2 domain: structural basis and relevance to the XLP syndrome. EMBO J 2002;21:314–323.
29. Lee CH, Kominos D, Jacques S, Margolis B, Schlessinger J, Shoelson SE, Kuriyan J. Crystal structures of peptide complexes of the

amino-terminal SH2 domain of the Syp tyrosine phosphatase. Structure 1994;2:423–438.

30. Poy F, Yaffe MB, Sayos J, Saxena K, Morra M, Sumegi J, Cantley LC, Terhorst C, Eck MJ. Crystal structures of the XLP protein SAP reveal a class of SH2 domains with extended, phosphotyrosine-independent sequence recognition. Mol Cell 1999;4:555–561.
31. Siegal G, Davis B, Kristensen SM, Sankar A, Linacre J, Stein RC, Panayotou G, Waterfield MD, Driscoll PC. Solution structure of the C-terminal SH2 domain of the p85 α regulatory subunit of phosphoinositide 3-kinase. J Mol Biol 1998;276:461–478.
32. Weber T, Schaffhausen B, Liu Y, Gunther UL. NMR structure of the N-SH2 of the p85 subunit of phosphoinositide 3-kinase complexed to a doubly phosphorylated peptide reveals a second phosphotyrosine binding site. Biochemistry 2000;39:15860–15869.
33. Huang KC, Cheng HT, Pai MT, Tzeng SR, Cheng JW. Solution structure and phosphopeptide binding of the SH2 domain from the human Bruton's tyrosine kinase. J Biomol NMR 2006;36:73–78.
34. Hof P, Pluskey S, Dhe-Paganon S, Eck MJ, Shoelson SE. Crystal structure of the tyrosine phosphatase SHP-2. Cell 1998;92:441–450.
35. Piirilä H, Väliaho J, Vihinen M. Immunodeficiency mutation databases (IDbases). Hum Mutat 2006;27:1200–1208.
36. Väliaho J, Smith CIE, Vihinen M. BTKbase: the mutation database for X-linked agammaglobulinemia. Hum Mutat 2006;27:1209–1217.
37. Lappalainen I, Giliani S, Franceschini R, Bonnefoy JY, Duckett C, Notarangelo LD, Vihinen M. Structural basis for SH2D1A mutations in X-linked lymphoproliferative disease. Biochem Biophys Res Commun 2000;269:124–130.
38. Vihinen M, Kwan SP, Lester T, Ochs HD, Resnick I, Väliaho J, Conley ME, Smith CIE. Mutations of the human BTK gene coding for bruton tyrosine kinase in X-linked agammaglobulinemia. Hum Mutat 1999;13:280–285.
39. Riikonen P, Vihinen M. MUTbase: maintenance and analysis of distributed mutation databases. Bioinformatics 1999;15:852–859.
40. Finn RD, Mistry J, Schuster-Bockler B, Griffiths-Jones S, Hollich V, Lassmann T, Moxon S, Marshall M, Khanna A, Durbin R, Eddy SR, Sonnhammer EL, Bateman A. Pfam: clans, web tools and services. Nucleic Acids Res 2006;34(Database issue):D247–D251.
41. O'Sullivan O, Suhre K, Abergel C, Higgins DG, Notredame C. 3DCoffee: combining protein sequences and structures within multiple sequence alignments. J Mol Biol 2004;340:385–395.
42. Frishman D, Argos P. Knowledge-based protein secondary structure assignment. Proteins 1995;23:566–579.
43. Shen B, Vihinen M. 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 2004;17:267–276.
44. Berezin C, Glaser F, Rosenberg J, Paz I, Pupko T, Fariselli P, Casadio R, Ben-Tal N. ConSeq: the identification of functionally and structurally important residues in protein sequences. Bioinformatics 2004;20:1322–1324.
45. Nolte RT, Eck MJ, Schlessinger J, Shoelson SE, Harrison SC. Crystal structure of the PI 3-kinase p85 amino-terminal SH2 domain and its phosphopeptide complexes. Nat Struct Biol 1996;3:364–374.
46. Mao X, Ren Z, Parker GN, Sondermann H, Pastorello MA, Wang W, McMurray JS, Demeler B, Darnell JE, Jr, Chen X. Structural bases of unphosphorylated STAT1 association and receptor binding. Mol Cell 2005;17:761–771.
47. Word JM, Bateman RC, Jr, Presley BK, Lovell SC, Richardson DC. Exploring steric constraints on protein mutations using MAGE/PROBE. Protein Sci 2000;9:2251–2259.
48. Lovell SC, Word JM, Richardson JS, Richardson DC. The penultimate rotamer library. Proteins 2000;40:389–408.
49. Word JM, Lovell SC, LaBean TH, Taylor HC, Zalis ME, Presley BK, Richardson JS, Richardson DC. Visualizing and quantifying molecular goodness-of-fit: small-probe contact dots with explicit hydrogen atoms. J Mol Biol 1999;285:1711–1733.
50. DeLano W. The PyMOL molecular graphics system. Palo Alto, CA: DeLano Scientific; 2002.
51. Vriend G. WHAT IF: a molecular modeling and drug design program. J Mol Graph 1990;8:52–56, 29.
52. Shen B, Vihinen M. RankViaContact: ranking and visualization of amino acid contacts. Bioinformatics 2003;19:2161–2162.
53. Sobolev V, Sorokine A, Prilusky J, Abola EE, Edelman M. Automated analysis of interatomic contacts in proteins. Bioinformatics 1999;15:327–332.
54. Thusberg J, Vihinen M. Bioinformatic analysis of protein structure-function relationships: case study of leukocyte elastase (ELA2) missense mutations. Hum Mutat 2006;27:1230–1243.
55. Thusberg J, Vihinen M. The structural basis of hyper IgM deficiency—CD40L mutations. Protein Eng Des Sel 2007;20:133–141.
56. Lappalainen I, Vihinen M. Structural basis of ICF-causing mutations in the methyltransferase domain of DNMT3B. Protein Eng 2002;15:1005–1014.
57. Rong SB, Vihinen M. Structural basis of Wiskott-Aldrich syndrome causing mutations in the WH1 domain. J Mol Med 2000;78:530–537.
58. Rong SB, Väliaho J, Vihinen M. Structural basis of Bloom syndrome (BS) causing mutations in the BLM helicase domain. Mol Med 2000;6:155–164.
59. Stenson PD, Ball EV, Mort M, Phillips AD, Shiel JA, Thomas NS, Abeysinghe S, Krawczak M, Cooper DN. Human Gene Mutation Database (HGMD): 2003 update. Hum Mutat 2003;21:577–581.
60. Strickler SS, Gribenko AV, Keiffer TR, Tomlinson J, Reihle T, Loladze VV, Makhatadze GI. Protein stability and surface electrostatics: a charged relationship. Biochemistry 2006;45:2761–2766.
61. Eck MJ, Shoelson SE, Harrison SC. Recognition of a high-affinity phosphotyrosyl peptide by the Src homology-2 domain of p56lck. Nature 1993;362:87–91.
62. Kuriyan J, Cowburn D. Modular peptide recognition domains in eukaryotic signaling. Annu Rev Biophys Biomol Struct 1997;26:259–288.
63. Tartaglia M, Mehler EL, Goldberg R, Zampino G, Brunner HG, Kremer H, van der Burgt I, Crosby AH, Ion A, Jeffery S, Kalidas K, Patton MA, Kucherlapati RS, Gelb BD. Mutations in PTPN11, encoding the protein tyrosine phosphatase SHP-2, cause Noonan syndrome. Nat Genet 2001;29:465–468.
64. Tartaglia M, Martinelli S, Stella L, Bocchinfuso G, Flex E, Cordeddu V, Zampino G, Burgt I, Palleschi A, Petrucci TC, Sorcini M, Schoch C, Foa R, Emanuel PD, Gelb BD. Diversity and functional consequences of germline and somatic PTPN11 mutations in human disease. Am J Hum Genet 2006;78:279–290.
65. Li C, Iosef C, Jia CY, Gkourasas T, Han VK, Shun-Cheng Li S. Disease-causing SAP mutants are defective in ligand binding and protein folding. Biochemistry 2003;42:14885–14892.
66. Hare NJ, Ma CS, Alvaro F, Nichols KE, Tangye SG. Missense mutations in SH2D1A identified in patients with X-linked lymphoproliferative disease differentially affect the expression and function of SAP. Int Immunol 2006;18:1055–1065.
67. Erdös M, Uzvolgyi E, Nemes Z, Torok O, Rakoczi E, Went-Sumegi N, Sumegi J, Marodi L. Characterization of a new disease-causing mutation of SH2D1A in a family with X-linked lymphoproliferative disease. Hum Mutat 2005;25:506.
68. Randles LG, Lappalainen I, Fowler SB, Moore B, Hamill SJ, Clarke J. Using model proteins to quantify the effects of pathogenic mutations in IG-like proteins. J Biol Chem 2006;34:24216–24226.
69. Matsuda S, Suzuki-Fujimoto T, Minowa A, Ueno H, Katamura K, Koyasu S. Temperature-sensitive ZAP70 mutants degrading through a proteasome-independent pathway. Restoration of a kinase domain mutant by Cdc37. J Biol Chem 1999;274:34515–34518.
70. Dupuis S, Jouanguy E, Al-Hajjar S, Fieschi C, Al-Mohsen IZ, Al-Jumaah S, Yang K, Chapgier A, Eidenschenk C, Eid P, Al Ghonaium A, Tufenkeji H, Frayha H, Al-Gazlan S, Al-Rayes H, Schreiber RD, Gresser I, Casanova JL. Impaired response to interferon-α/β and lethal viral disease in human STAT1 deficiency. Nat Genet 2003;33:388–391.

71. Chia DJ, Subbian E, Buck TM, Hwa V, Rosenfeld RG, Skach WR, Shinde U, Rotwein P. Aberrant folding of a mutant Stat5b causes growth hormone insensitivity and proteasomal dysfunction. J Biol Chem 2006;281:6552–6558.
72. Poole AW, Jones ML. A SHPing tale: perspectives on the regulation of SHP-1 and SHP-2 tyrosine phosphatases by the C-terminal tail. Cell Signal 2005;17:1323–1332.
73. Tzeng SR, Pai MT, Lung FD, Wu CW, Roller PP, Lei B, Wei CJ, Tu SC, Chen SH, Soong WJ, Cheng JW. Stability and peptide binding specificity of Btk SH2 domain: molecular basis for X-linked agammaglobulinemia. Protein Sci 2000;9:2377–2385.
74. Vihinen M, Arredondo-Vega FX, Casanova JL, Etzioni A, Giliani S, Hammarström L, Hershfield MS, Heyworth PG, Hsu AP, Lähdesmäki A, Lappalainen I, Notarangelo LD, Puck JM, Reith W, Roos D, Schumacher RF, Schwarz K, Vezzoni P, Villa A, Väliaho J, Smith CIE. Primary immunodeficiency mutation databases. Adv Genet 2001;43:103–188.
75. Warnecke PM, Bestor TH. Cytosine methylation and human cancer. Curr Opin Oncol 2000;12:68–73.
76. Baraldi E, Carugo KD, Hyvonen M, Surdo PL, Riley AM, Potter BV, O'Brien R, Ladbury JE, Saraste M. Structure of the PH domain from Bruton's tyrosine kinase in complex with inositol 1,3,4,5-tetrakisphosphate. Structure 1999;7:449–460.
77. Bullock AN, Fersht AR. Rescuing the function of mutant p53. Nat Rev Cancer 2001;1:68–76.
78. Partridge AW, Therien AG, Deber CM. Missense mutations in transmembrane domains of proteins: phenotypic propensity of polar residues for human disease. Proteins 2004;54:648–656.
79. Bradshaw JM, Mitaxov V, Waksman G. Investigation of phosphotyrosine recognition by the SH2 domain of the Src kinase. J Mol Biol 1999;293:971–985.
80. Henriques DA, Ladbury JE, Jackson RM. Comparison of binding energies of SrcSH2-phosphotyrosyl peptides with structure-based prediction using surface area based empirical parameterization. Protein Sci 2000;9:1975–1985.
81. Vořechovský I, Luo L, Hertz JM, Froland SS, Klemola T, Fiorini M, Quinti I, Paganelli R, Ozsahin H, Hammarström L, Webster AD, Smith CIE. Mutation pattern in the Bruton's tyrosine kinase gene in 26 unrelated patients with X-linked agammaglobulinemia. Hum Mutat 1997; 9:418–425.
82. Vassilev A, Ozer Z, Navara C, Mahajan S, Uckun FM. Bruton's tyrosine kinase as an inhibitor of the Fas/CD95 death-inducing signaling complex. J Biol Chem 1999;274:1646–1656.
83. Songyang Z, Shoelson SE, Chaudhuri M, Gish G, Pawson T, Haser WG, King F, Roberts T, Ratnofsky S, Lechleider RJ, Neel BG, Birge RB, Fajardo JE, Chou MM, Hanafusa H, Schaffhausen B, Cantley LC. SH2 domains recognize specific phosphopeptide sequences. Cell 1993;72:767–778.
84. Vihinen M, Durandy A. Primary immunodeficiencies: phenotype genotype correlations. In: Falus A, editor. Immunogenomics and human disease. West Sussex: Wiley; 2005. pp 443–460.
85. Kratz CP, Niemeyer CM, Castleberry RP, Cetin M, Bergstrasser E, Emanuel PD, Hasle H, Kardos G, Klein C, Kojima S, Stary J, Trebo M, Zecca M, Gelb BD, Tartaglia M, Loh ML. The mutational spectrum of PTPN11 in juvenile myelomonocytic leukemia and Noonan syndrome/myeloproliferative disease. Blood 2005;106:2183–2185.
86. Kofoed EM, Hwa V, Little B, Woods KA, Buckway CK, Tsubaki J, Pratt KL, Bezrodnik L, Jasper H, Tepper A, Heinrich JJ, Rosenfeld RG. Growth hormone insensitivity associated with a STAT5b mutation. N Engl J Med 2003;349:1139–1147.

SUPPLEMENTARY TEXT

GENOME WIDE ANALYSIS OF PATHOGENIC SH2 DOMAIN MUTATIONS

Ilkka Lappalainen, Janita Thusberg, Bairong Shen, and Mauno Vihinen

Biological processes of the disease-related SH2 domain-containing proteins

BTK and ZAP-70 proteins are cytoplasmic tyrosine kinases having a profound role in B and T cell maturation, respectively. Mutations in all five domains of BTK have been shown to block B cell maturation after pre-B-cell causing X-linked agammaglobulinemia (XLA).[1] BTK participates in signal transduction pathways regulating activation, proliferation, differentiation, and apoptosis, initiated by the binding of a variety of extracellular ligands to cell surface receptors.[2] Mutations in the SH2 domain of BTK have been shown to abrogate sustained Ca^{2+} mobilization and disrupt the binding to the Fas receptor.[3] The ZAP-70 protein consists of two SH2 domains followed by a C-terminal kinase domain[4] (Figure 1). Association of both SH2 domains to the ζ chain of activated T cell antigen receptor (TCR) regulates multiple different downstream pathways.[4] Alterations in the *ZAP70* gene lead to a rare autosomal recessive form of severe combined immunodeficiency (SCID).[5]

The gene defective in X-linked lymphoproliferative syndrome (XLP) encodes a protein, SH2D1A, or SAP, expressed mainly in the T cells and natural killer (NK) cells.[6-8] The SH2D1A protein belongs to a family of small regulator proteins together with Ewing’s sarcoma-activated transcript 2 (EAT-2) and EAT-2-related transducer (ERT). ERT is functional in rodents, but the human ERT is a pseudogene.[9] The three proteins consist of a single SH2 domain and a C-terminal extension[9,10] (Figure 1). According to the mutational and structural data, the function of the SH2D1A protein seems to compete with SHP-1 and SHP-2 for binding to the same consensus motifs in the cytoplasmic tail of several members of the immunoglobulin superfamily, including SLAM (signal lymphocyte-activator molecule), 2B4, CD84, and Ly-9.[8,11-13] These receptors have been shown to function as activators and adhesion molecules in the immune synapse between the T and NK cells and the antigen presenting cells. SH2D1A affects downstream signalling in several ways and therefore its dysfunction leads to the broad clinical spectrum of XLP.

The *PTPN11* gene encodes SHP-2, a ubiquitously expressed cytoplasmic tyrosine phosphatase that consists of two tandemly arranged SH2 domains at the N-terminus, a catalytic domain, and a C-terminal tail containing a proline-rich region and two tyrosyl residues that undergo reversible phosphorylation[14] (Figure 1). SHP-2 is a critical component in several signalling pathways involved in the control of developmental processes,[15-19] hematopoiesis,[18,20,21] and metabolism.[22] Mutations in *PTPN11* cause Noonan syndrome (NS), a developmental disorder characterized by facial dysmorphisms, short stature, skeletal and haematological defects, and cardiovascular abnormalities.[23,24] Noonan syndrome is also caused by mutations in GTPase KRas (*KRAS*),[25] and son of sevenless homolog 1 (*SOS-1*)[26] which does not contain an SH2 domain. Leopard sydrome (LS),[27,28] a clinically related disorder, is caused by mutations in the SHP-2 PTP domain. *PTPN11* mutations also occur in several human cancers, including juvenile myelomonocytic leukaemia (JMML), myelodysplastic syndrome (MDS), B-cell acute lymphoblastic leukaemia (BLL), and acute myelogeneous leukaemia (AML).[29] The activating *PTPN11* mutations play a broad role in cancer, because SHP-2 acts as a signal-enhancing signalling component in pathways that regulate cell growth, transformation, differentiation, and migration. The protein is also required for normal Ras activation in many of these pathways.[14]

The class Ia PI3-kinase plays a pivotal role in signal transduction pathways linking insulin with many of its specific cellular responses, including GLUT4 vesicle translocation to the plasma membrane and the inhibition of glycogen synthase kinase-3.[30] Moreover, PI3-kinase is necessary, if not sufficient, for the insulin-stimulated increase in glucose uptake, and glycogen synthesis in insulin-sensitive tissues.[31,32] The PI3-kinase is heterodimeric, consisting of a catalytic subunit (p110), and a regulatory subunit (p85α).[33] A missense mutation has been found in the N-terminal SH2 domain (N-SH2) of p85α, leading to a severe insulin resistance.[34] The multiple genetic and environmental influences on insulin sensitivity are reflected by the fact that even in families with pathogenic insulin receptor mutations there is enormous inter-individual variability in the severity of hyperinsulinemia.[35]

Basal cell carcinoma (BCC) is the most frequent skin cancer in the white population.[36,37] BCCs mostly occur sporadically in relation to sun exposure, although their incidence is increased significantly in some rare genetic disorders.[38-40] Mutations within the C-terminal SH2 domain (C-SH2) of the GTPase activating protein RasGAP have also been found in a subset of BCCs. RasGAP acts by enhancing the intrinsic GTPase activity of Ras, leading to the hydrolysis of bound GTP to GDP and down-regulation of Ras activity.[41-43] The region in which the mutations are clustered is A/T rich; raising the possibility that UV radiation may be a contributing factor in the development of the tumours.[44]

The STAT SH2 domains are divergent in sequence from most other SH2 domains,[45,46] but the basic architecture and the mechanism for recognizing the phosphotyrosyl polypeptide are both fundamentally the same as that elucidated for other SH2 domains.[47] The STAT SH2 domains act as phosphorylation-dependent switches that control receptor recognition and DNA binding.[48] Upon cytokine and growth factor stimulation of cells, STAT molecules become tyrosine phosphorylated, dimerize through reciprocal phosphotyrosine-SH2 interactions, accumulate in the nucleus, bind to DNA, and activate gene transcription.[49]

STAT1 mediates interferon signalling as a part of the JAK-STAT1-pathway. After being phosphorylated, STAT1 undergoes dimerization through its SH2 domains, translocates to the nucleus and regulates gene expression by binding to γ-activated sequence elements in the promoters of IFN-γ regulated genes.[48,50,51] STAT1 knockout mice develop normally, but lack the classical responses to IFNγ and IFN α/β, and are therefore extremely susceptible to viral and bacterial infections,[52,53] which illustrates the critical role of STAT1 in the function of the immune system. The complete STAT1 deficiency causes impaired response to IFN-γ, leading to severe viral disease and mycobacteriosis. The deficiency has been described only in two unrelated patients.[54] STAT5B acts as a part of the growth hormone signalling pathway leading to stimulation of insulin-like growth factor I (IGF-I) gene transcription.[55] The absence of STAT5B is associated with diminished post-natal growth, as demonstrated by mouse knockout models,[56,57] and defects in the STAT5B SH2 domain in humans lead to growth hormone insensitivity with immunodeficiency.[58]

References

1. Väliaho J, Smith CIE, Vihinen M. BTKbase: the mutation database for X-linked agammaglobulinemia. Hum Mutat 2006;27(12):1209-1217.
2. Lindvall JM, Blomberg KE, Väliaho J, Vargas L, Heinonen JE, Berglöf A, Mohamed AJ, Nore BF, Vihinen M, Smith CIE. Bruton's tyrosine kinase: cell biology, sequence conservation, mutation spectrum, siRNA modifications, and expression profiling. Immunol Rev 2005;203:200-215.

3. Vassilev A, Ozer Z, Navara C, Mahajan S, Uckun FM. Bruton's tyrosine kinase as an inhibitor of the Fas/CD95 death-inducing signaling complex. J Biol Chem 1999;274(3):1646-1656.

4. Chan AC, Iwashima M, Turck CW, Weiss A. ZAP-70: a 70 kd protein-tyrosine kinase that associates with the TCR zeta chain. Cell 1992;71(4):649-662.

5. Chan AC, Kadlecek TA, Elder ME, Filipovich AH, Kuo WL, Iwashima M, Parslow TG, Weiss A. ZAP-70 deficiency in an autosomal recessive form of severe combined immunodeficiency. Science 1994;264(5165):1599-1601.

6. Coffey AJ, Brooksbank RA, Brandau O, Oohashi T, Howell GR, Bye JM, Cahn AP, Durham J, Heath P, Wray P, Pavitt R, Wilkinson J, Leversha M, Huckle E, Shaw-Smith CJ, Dunham A, Rhodes S, Schuster V, Porta G, Yin L, Serafini P, Sylla B, Zollo M, Franco B, Bolino A, Seri M, Lanyi A, Davis JR, Webster D, Harris A, Lenoir G, de St Basile G, Jones A, Behloradsky BH, Achatz H, Murken J, Fassler R, Sumegi J, Romeo G, Vaudin M, Ross MT, Meindl A, Bentley DR. Host response to EBV infection in X-linked lymphoproliferative disease results from mutations in an SH2-domain encoding gene. Nat Genet 1998;20(2):129-135.

7. Nichols KE, Harkin DP, Levitz S, Krainer M, Kolquist KA, Genovese C, Bernard A, Ferguson M, Zuo L, Snyder E, Buckler AJ, Wise C, Ashley J, Lovett M, Valentine MB, Look AT, Gerald W, Housman DE, Haber DA. Inactivating mutations in an SH2 domain-encoding gene in X-linked lymphoproliferative syndrome. Proc Natl Acad Sci USA 1998;95(23):13765-13770.

8. Sayos J, Wu C, Morra M, Wang N, Zhang X, Allen D, van Schaik S, Notarangelo L, Geha R, Roncarolo MG, Oettgen H, De Vries JE, Aversa G, Terhorst C. The X-linked lymphoproliferative-disease gene product SAP regulates signals induced through the co-receptor SLAM. Nature 1998;395(6701):462-469.

9. Roncagalli R, Taylor JE, Zhang S, Shi X, Chen R, Cruz-Munoz ME, Yin L, Latour S, Veillette A. Negative regulation of natural killer cell function by EAT-2, a SAP-related adaptor. Nat Immunol 2005;6(10):1002-1010.

10. Morra M, Lu J, Poy F, Martin M, Sayos J, Calpe S, Gullo C, Howie D, Rietdijk S, Thompson A, Coyle AJ, Denny C, Yaffe MB, Engel P, Eck MJ, Terhorst C. Structural basis for the interaction of the free SH2 domain EAT-2 with SLAM receptors in hematopoietic cells. Embo J 2001;20(21):5840-5852.

11. Tangye SG, Lazetic S, Woollatt E, Sutherland GR, Lanier LL, Phillips JH. Cutting edge: human 2B4, an activating NK cell receptor, recruits the protein tyrosine phosphatase SHP-2 and the adaptor signaling protein SAP. J Immunol 1999;162(12):6981-6985.

12. Lewis J, Eiben LJ, Nelson DL, Cohen JI, Nichols KE, Ochs HD, Notarangelo LD, Duckett CS. Distinct interactions of the X-linked lymphoproliferative syndrome gene product SAP with cytoplasmic domains of members of the CD2 receptor family. Clin Immunol 2001;100(1):15-23.

13. Sayos J, Martin M, Chen A, Simarro M, Howie D, Morra M, Engel P, Terhorst C. Cell surface receptors Ly-9 and CD84 recruit the X-linked lymphoproliferative disease gene product SAP. Blood 2001;97(12):3867-3874.

14. Neel BG, Gu H, Pao L. The 'Shp'ing news: SH2 domain-containing tyrosine phosphatases in cell signaling. Trends Biochem Sci 2003;28(6):284-293.

15. Tang TL, Freeman RM, Jr., O'Reilly AM, Neel BG, Sokol SY. The SH2-containing protein-tyrosine phosphatase SH-PTP2 is required upstream of MAP kinase for early Xenopus development. Cell 1995;80(3):473-483.

16. Saxton TM, Henkemeyer M, Gasca S, Shen R, Rossi DJ, Shalaby F, Feng GS, Pawson T. Abnormal mesoderm patterning in mouse embryos mutant for the SH2 tyrosine phosphatase Shp-2. Embo J 1997;16(9):2352-2364.

17. Saxton TM, Ciruna BG, Holmyard D, Kulkarni S, Harpal K, Rossant J, Pawson T. The SH2 tyrosine phosphatase shp2 is required for mammalian limb development. Nat Genet 2000;24(4):420-423.

18. Qu CK, Yu WM, Azzarelli B, Cooper S, Broxmeyer HE, Feng GS. Biased suppression of hematopoiesis and multiple developmental defects in chimeric mice containing Shp-2 mutant cells. Mol Cell Biol 1998;18(10):6075-6082.

19. Chen B, Bronson RT, Klaman LD, Hampton TG, Wang JF, Green PJ, Magnuson T, Douglas PS, Morgan JP, Neel BG. Mice mutant for Egfr and Shp2 have defective cardiac semilunar valvulogenesis. Nat Genet 2000;24(3):296-299.

20. Qu CK, Shi ZQ, Shen R, Tsai FY, Orkin SH, Feng GS. A deletion mutation in the SH2-N domain of Shp-2 severely suppresses hematopoietic cell development. Mol Cell Biol 1997;17(9):5499-5507.

21. Qu CK, Nguyen S, Chen J, Feng GS. Requirement of Shp-2 tyrosine phosphatase in lymphoid and hematopoietic cell development. Blood 2001;97(4):911-914.

22. Zhang EE, Chapeau E, Hagihara K, Feng GS. Neuronal Shp2 tyrosine phosphatase controls energy balance and metabolism. Proc Natl Acad Sci USA 2004;101(45):16064-16069.

23. Noonan JA. Hypertelorism with Turner phenotype. A new syndrome with associated congenital heart disease. Am J Dis Child 1968;116(4):373-380.

24. Allanson JE. Noonan syndrome. J Med Genet 1987;24(1):9-13.

25. Schubbert S, Zenker M, Rowe SL, Boll S, Klein C, Bollag G, van der Burgt I, Musante L, Kalscheuer V, Wehner LE, Nguyen H, West B, Zhang KY, Sistermans E, Rauch A, Niemeyer CM, Shannon K, Kratz CP. Germline KRAS mutations cause Noonan syndrome. Nat Genet 2006;38(3):331-336.

26. Roberts AE, Araki T, Swanson KD, Montgomery KT, Schiripo TA, Joshi VA, Li L, Yassin Y, Tamburino AM, Neel BG, Kucherlapati RS. Germline gain-of-function mutations in SOS1 cause Noonan syndrome. Nat Genet 2007;39(1):70-74.

27. Digilio MC, Conti E, Sarkozy A, Mingarelli R, Dottorini T, Marino B, Pizzuti A, Dallapiccola B. Grouping of multiple-lentigines/LEOPARD and Noonan syndromes on the PTPN11 gene. Am J Hum Genet 2002;71(2):389-394.

28. Legius E, Schrander-Stumpel C, Schollen E, Pulles-Heintzberger C, Gewillig M, Fryns JP. PTPN11 mutations in LEOPARD syndrome. J Med Genet 2002;39(8):571-574.

29. Bentires-Alj M, Paez JG, David FS, Keilhack H, Halmos B, Naoki K, Maris JM, Richardson A, Bardelli A, Sugarbaker DJ, Richards WG, Du J, Girard L, Minna JD, Loh ML, Fisher DE, Velculescu VE, Vogelstein B, Meyerson M, Sellers WR, Neel BG. Activating mutations of the noonan syndrome-associated SHP2/PTPN11 gene in human solid tumors and adult acute myelogenous leukemia. Cancer Res 2004;64(24):8816-8820.

30. Cohen P. The twentieth century struggle to decipher insulin signalling. Nat Rev Mol Cell Biol 2006;7(11):867-873.

31. Shepherd PR, Withers DJ, Siddle K. Phosphoinositide 3-kinase: the key switch mechanism in insulin signalling. Biochem J 1998;333 (Pt 3):471-490.

32. Holman GD, Kasuga M. From receptor to transporter: insulin signalling to glucose transport. Diabetologia 1997;40(9):991-1003.

33. Antonetti DA, Algenstaedt P, Kahn CR. Insulin receptor substrate 1 binds two novel splice variants of the regulatory subunit of phosphatidylinositol 3-kinase in muscle and brain. Mol Cell Biol 1996;16(5):2195-2203.

34. Baynes KC, Beeton CA, Panayotou G, Stein R, Soos M, Hansen T, Simpson H, O'Rahilly S, Shepherd PR, Whitehead JP. Natural variants of human p85 alpha phosphoinositide 3-kinase

in severe insulin resistance: a novel variant with impaired insulin-stimulated lipid kinase activity. Diabetologia 2000;43(3):321-331.

35. Moller DE, Flier JS. Insulin resistance--mechanisms, syndromes, and implications. N Engl J Med 1991;325(13):938-948.

36. Miller SJ. Biology of basal cell carcinoma (Part I). J Am Acad Dermatol 1991;24(1):1-13.

37. Miller SJ. Biology of basal cell carcinoma (Part II). J Am Acad Dermatol 1991;24(2 Pt 1):161-175.

38. Gorlin RJ. Nevoid basal-cell carcinoma syndrome. Medicine (Baltimore) 1987;66(2):98-113.

39. Goeteyn M, Geerts ML, Kint A, De Weert J. The Bazex-Dupre-Christol syndrome. Arch Dermatol 1994;130(3):337-342.

40. Bodak N, Queille S, Avril MF, Bouadjar B, Drougard C, Sarasin A, Daya-Grosjean L. High levels of patched gene mutations in basal-cell carcinomas from patients with xeroderma pigmentosum. Proc Natl Acad Sci USA 1999;96(9):5117-5122.

41. Gold MR, Crowley MT, Martin GA, McCormick F, DeFranco AL. Targets of B lymphocyte antigen receptor signal transduction include the p21ras GTPase-activating protein (GAP) and two GAP-associated proteins. J Immunol 1993;150(2):377-386.

42. Lazarus AH, Kawauchi K, Rapoport MJ, Delovitch TL. Antigen-induced B lymphocyte activation involves the p21ras and ras.GAP signaling pathway. J Exp Med 1993;178(5):1765-1769.

43. Scheffzek K, Ahmadian MR, Wittinghofer A. GTPase-activating proteins: helping hands to complement an active site. Trends Biochem Sci 1998;23(7):257-262.

44. Friedman E, Gejman PV, Martin GA, McCormick F. Nonsense mutations in the C-terminal SH2 region of the GTPase activating protein (GAP) gene in human tumours. Nat Genet 1993;5(3):242-247.

45. Darnell JE, Jr. Phosphotyrosine signaling and the single cell:metazoan boundary. Proc Natl Acad Sci USA 1997;94(22):11767-11769.

46. Kawata T, Shevchenko A, Fukuzawa M, Jermyn KA, Totty NF, Zhukovskaya NV, Sterling AE, Mann M, Williams JG. SH2 signaling in a lower eukaryote: a STAT protein that regulates stalk cell differentiation in dictyostelium. Cell 1997;89(6):909-916.

47. Kuriyan J, Cowburn D. Modular peptide recognition domains in eukaryotic signaling. Annu Rev Biophys Biomol Struct 1997;26:259-288.

48. Chen X, Vinkemeier U, Zhao Y, Jeruzalmi D, Darnell JE, Jr., Kuriyan J. Crystal structure of a tyrosine phosphorylated STAT-1 dimer bound to DNA. Cell 1998;93(5):827-839.

49. Mao X, Ren Z, Parker GN, Sondermann H, Pastorello MA, Wang W, McMurray JS, Demeler B, Darnell JE, Jr., Chen X. Structural bases of unphosphorylated STAT1 association and receptor binding. Mol Cell 2005;17(6):761-771.

50. Darnell JE, Jr. STATs and gene regulation. Science 1997;277(5332):1630-1635.

51. Ramana CV, Chatterjee-Kishore M, Nguyen H, Stark GR. Complex roles of Stat1 in regulating gene expression. Oncogene 2000;19(21):2619-2627.

52. Meraz MA, White JM, Sheehan KC, Bach EA, Rodig SJ, Dighe AS, Kaplan DH, Riley JK, Greenlund AC, Campbell D, Carver-Moore K, DuBois RN, Clark R, Aguet M, Schreiber RD. Targeted disruption of the Stat1 gene in mice reveals unexpected physiologic specificity in the JAK-STAT signaling pathway. Cell 1996;84(3):431-442.

53. Durbin JE, Hackenmiller R, Simon MC, Levy DE. Targeted disruption of the mouse Stat1 gene results in compromised innate immunity to viral disease. Cell 1996;84(3):443-450.

54. Dupuis S, Jouanguy E, Al-Hajjar S, Fieschi C, Al-Mohsen IZ, Al-Jumaah S, Yang K, Chapgier A, Eidenschenk C, Eid P, Al Ghonaium A, Tufenkeji H, Frayha H, Al-Gazlan S, Al-Rayes H, Schreiber RD, Gresser I, Casanova JL. Impaired response to interferon-

alpha/beta and lethal viral disease in human STAT1 deficiency. Nat Genet 2003;33(3):388-391.

55. Woelfle J, Billiard J, Rotwein P. Acute control of insulin-like growth factor-I gene transcription by growth hormone through Stat5b. J Biol Chem 2003;278(25):22696-22702.

56. Udy GB, Towers RP, Snell RG, Wilkins RJ, Park SH, Ram PA, Waxman DJ, Davey HW. Requirement of STAT5b for sexual dimorphism of body growth rates and liver gene expression. Proc Natl Acad Sci USA 1997;94(14):7239-7244.

57. Teglund S, McKay C, Schuetz E, van Deursen JM, Stravopodis D, Wang D, Brown M, Bodner S, Grosveld G, Ihle JN. Stat5a and Stat5b proteins have essential and nonessential, or redundant, roles in cytokine responses. Cell 1998;93(5):841-850.

58. Kofoed EM, Hwa V, Little B, Woods KA, Buckway CK, Tsubaki J, Pratt KL, Bezrodnik L, Jasper H, Tepper A, Heinrich JJ, Rosenfeld RG. Growth hormone insensitivity associated with a STAT5b mutation. N Engl J Med 2003;349(12):1139-1147.

REVIEW

Human Mutation

OFFICIAL JOURNAL HGVS HUMAN GENOME VARIATION SOCIETY www.hgvs.org

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

Janita Thusberg[1] and Mauno Vihinen[1,2]*

[1]Institute of Medical Technology, FI-33014 University of Tampere, Finland; [2]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.

© 2009 WILEY-LISS, INC.

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 Moult, 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].

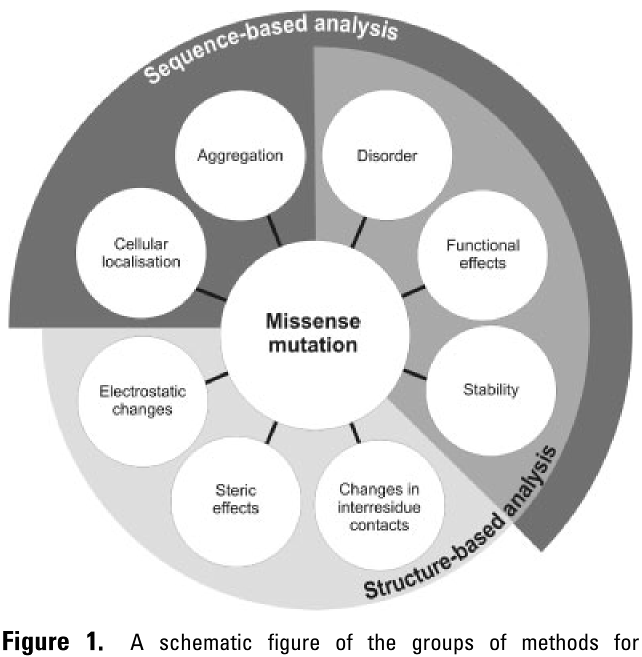


Figure 1. A schematic figure of the groups of methods for analyzing the effects of missense mutations. Our approach can be divided into sequence- and structure-based sections (dark gray and light gray backgrounds, respectively), which in part overlap.

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 (CMDBs), 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 CMDBs (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., hydropathy, 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.

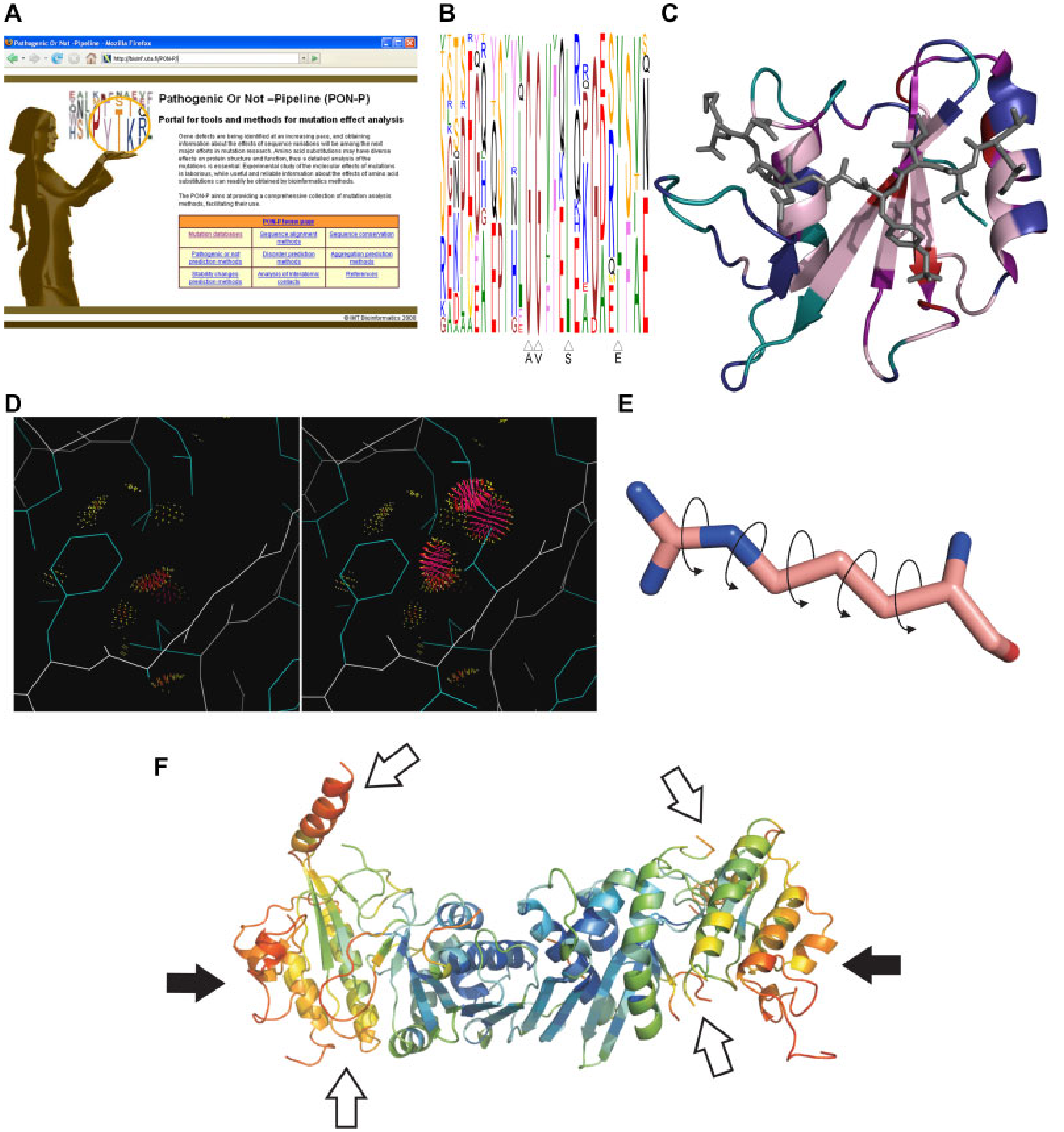
Protein Localization

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 [Tompa, 2002] and chaperone [Tompa 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



Pathogenic Or Not -Pipeline - Mozilla Firefox

http://bioinf.uta.fi/PON-P

Pathogenic Or Not –Pipeline (PON-P)

Portal for tools and methods for mutation effect analysis

Gene defects are being identified at an increasing pace, and obtaining information about the effects of sequence variations 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, while useful and reliable information about the effects of amino acid substitutions can readily be obtained by bioinformatics methods.

The PON-P aims at providing a comprehensive collection of mutation analysis methods, facilitating their use.

PON-P home page		
Mutation databases	Sequence alignment methods	Sequence conservation
Pathogenic or not prediction methods	Disorder prediction methods	Aggregation prediction methods
Stability changes prediction methods	Analysis of interatomic contacts	References

© IMT Bioinformatics 2008

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 ≥0.25 Å; hot pink—serious clash overlap ≥0.4 Å. The figure is created by KiNG [Lovell et al., 2003]. **E**: Schematic representation of amino acid side chain χ 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 IIβ phosphatidylinositol phosphate kinase (PDB ID 1BO1) 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	http://snpanalyzer.utmem.edu/	Pathogenic or not	(Bao et al., 2005)
Panther	http://www.pantherdb.org/tools/csnpScoreForm.jsp	Conservation analysis, pathogenic or not	(Thomas et al., 2003)
PhD-SNP	http://gpcr.biocomp.unibo.it/cgi/predictors/PhD-SNP/PhD-SNP.cgi	Pathogenic or not	(Capriotti et al., 2006)
PMut	http://mmb2.pcb.ub.es:8080/PMut/	Pathogenic or not	(Ferrer-Costa et al., 2005)
PolyPhen	http://coot.embl.de/PolyPhen/	Pathogenic or not	(Ramensky et al., 2002)
SIFT	http://blocks.fhcrc.org/sift/SIFT.html	Pathogenic or not	(Ng and Henikoff, 2001)
SNAP	http://cubic.bioc.columbia.edu/services/SNAP/	Pathogenic or not	(Bromberg and Rost, 2007)
SNPs3D	http://www.snps3d.org/	Pathogenic or not	(Yue et al., 2006)
Sequence alignment methods			
M-Coffee	http://www.tcoffee.org/	Multiple sequence alignment	(Wallace et al., 2006)
MAFFT	http://align.bmr.kyushu-u.ac.jp/mafft/online/server/	Multiple sequence alignment	(Katoh et al., 2002, 2005)
PROBCONS	http://probcons.stanford.edu/	Multiple sequence alignment	(Do et al., 2005)
PROMALS	http://prodata.swmed.edu/promals/	Multiple sequence alignment	(Pei et al., 2007)
ClustalW2	http://www.ebi.ac.uk/Tools/clustalw2/index.html	Multiple sequence alignment	(Larkin et al., 2007)
MUSCLE	http://www.ebi.ac.uk/Tools/muscle/index.html	Multiple sequence alignment	(Edgar, 2004)
Conservation analysis			
ClustalX	http://www.ebi.ac.uk/Tools/clustalw2/index.html	Conservation analysis and visualization	(Larkin et al., 2007)
ConSeq	http://conseq.tau.ac.il/	Conservation analysis and visualization	(Berezin et al., 2004)
ConSSeq	http://sms.cbi.cnptia.embrapa.br/SMS/STINGm/consseq/	Conservation analysis and visualization	(Higa et al., 2004)
ConSurf	http://consurf.tau.ac.il/	Conservation analysis and visualization	(Glaser et al., 2003; Landau et al., 2005)
Jalview	http://www.jalview.org/	MSA visualization	(Clamp et al., 2004)
MatrixPlot	http://www.cbs.dtu.dk/services/MatrixPlot/	Conservation analysis and visualization	(Gorodkin et al., 1999)
MultiDisp	http://bioinf.uta.fi/cgi-bin/MultiDisp.cgi	Conservation analysis and visualization	(Riikonen and Vihinen, in preparation)
ProCon		Conservation analysis and visualization	(Shen and Vihinen, 2004)
Stability changes prediction			
Auto-Mute	http://proteins.gmu.edu/automute/AUTO-MUTE.html	Stability changes prediction	(Masso and Vaisman, 2008)
CUPSAT	http://cupsat.tu-bs.de/	Stability changes prediction	(Parthiban et al., 2006)
Dmutant	http://sparks.informatics.iupui.edu/hzhou/mutation.html	Stability changes prediction	(Zhou and Zhou, 2002)
Eris	http://troll.med.unc.edu/eris/login.php	Stability changes prediction	(Yin et al., 2007)
FoldX	http://foldx.crg.es/	Folding and stability changes prediction	(Guerois et al., 2002)
I-Mutant 2.0	http://gpcr2.biocomp.unibo.it/cgi/predictors/I-Mutant2.0/I-Mutant2.0.cgi	Stability changes prediction	(Capriotti, et al., 2005a, b)
MuPRO	http://www.ics.uci.edu/%7Ebaldig/mutation.html	Stability changes prediction	(Cheng et al., 2006)
PoPMuSiC	http://babylone.ulb.ac.be/popmusic/	Stability changes prediction	(Gilis and Rooman, 2000)
SCide	http://www.enzim.hu/scide/ide2.html	Stability changes prediction	(Dosztányi et al., 2003)
SCpred	http://www.enzim.hu/scpred/pred.html	Stability changes prediction	(Dosztányi et al., 1997)
SRide	http://sride.enzim.hu/	Stability changes prediction	(Magyar et al., 2005)
Disorder prediction			
CAST		Disorder prediction	(Promponas et al., 2000)
DisEMBL	http://dis.embl.de/	Disorder prediction	(Linding et al., 2003a)
Disopred	http://bioinf.cs.ucl.ac.uk/disopred/disopred.html	Disorder prediction	(Ward et al., 2004)
DISpro	http://scratch.proteomics.ics.uci.edu/	Disorder prediction	(Cheng et al., 2005)
Disprot	http://www.ist.temple.edu/disprot/predictor.php	Disorder prediction	(Obradović et al., 2003; Peng et al., 2005; Vucetic et al., 2003)
DRIPPRED	http://www.sbc.su.se/∼maccallr/disorder/	Disorder prediction	
FoldIndex	http://bip.weizmann.ac.il/fldbin/findex	Prediction of folding	(Prilusky et al., 2005)
FoldUnfold	http://skuld.protres.ru/∼mlobanov/ogu/ogu.cgi	Disorder prediction	(Galzitskaya et al., 2006)
GlobPlot	http://globplot.embl.de/	Disorder prediction	(Linding et al., 2003b)
iPDA	http://biominer.bime.ntu.edu.tw/ipda/	Disorder prediction	(Su et al., 2007)
IUPred	http://iupred.enzim.hu/	Disorder prediction	(Dosztányi et al., 2005)
NORSp	http://cubic.bioc.columbia.edu/services/NORSp/	Disorder prediction	(Liu and Rost, 2003)
PONDR	http://www.pondr.com/	Disorder prediction	(Obradović et al., 2005; Romero et al., 2001)
POODLE	http://mbs.cbrc.jp/poodle/poodle.html	Disorder prediction	(Hirose et al., 2007; Shimizu et al., 2007a, b)
PrDOS	http://prdos.hgc.jp	Disorder prediction	(Ishida and Kinoshita, 2007)
PreLink	http://genomics.eu.org/spip/PreLink	Disorder prediction	(Coeytaux and Poupon, 2005)
RONN	http://www.strubi.ox.ac.uk/RONN	Disorder prediction	(Yang et al., 2005)
SEG	http://mendel.imp.ac.at/METHODS/seg.server.html	Disorder prediction	(Wootton, 1994)
Spritz	http://protein.cribi.unipd.it/spritz/	Disorder prediction	(Vullo et al., 2006)
Analysis of interatomic contacts			
CMA	http://ligin.weizmann.ac.il/cma/	Analysis of interatomic contacts	(Sobolev et al., 2005)
CSU	http://bip.weizmann.ac.il/oca-bin/lpccsu	Analysis of interatomic contacts	(Sobolev et al., 1999)
KiNG	http://kinemage.biochem.duke.edu/software/king.php	Molecular graphics	(Lovell et al., 2003)
MolProbity	http://molprobity.biochem.duke.edu/	Analysis of interatomic contacts and packing, structure validation	(Davis et al., 2004)

Table 1. Continued

Service name	URL	Description	Reference
PROBE	http://kinemage.biochem.duke.edu/software/probe.php	Analysis of interatomic contacts and packing	(Word et al., 2000; Word et al., 1999)
PyMOL	http://pymol.sourceforge.net/	Molecular graphics	(DeLano, 2002)
RankViaContact	http://bioinf.uta.fi/RankViaContact.html	Analysis and visualization of interatomic contacts	(Shen and Vihinen, 2003)
Aggregation prediction			
Aggrescan	http://bioinf.uab.es/aggrescan/	Aggregation prediction	(Conchillo-Sole et al., 2007)
PASTA	http://protein.cribi.unipd.it/pasta/	Aggregation prediction	(Trovato et al., 2007)
TANGO	http://tango.embl.de/	Aggregation prediction	(Fernandez-Escamilla et al., 2004)
Waltz	http://switpc7.vub.ac.be/cgi-bin/submit.cgi	Aggregation prediction	(Maurer-Stroh et al., submitted for publication)
Other			
ExPASy Proteomics tools	http://ca.expasy.org/tools/#ptm	Posttranslational modification prediction tools	
SABLE	http://sable.cchmc.org/	Prediction of solvent accessibilities, 2D structures and transmembrane domains	(Adamczak et al., 2004, 2005; Wagner et al., 2005)
SNPeffect	http://snpeffect.vib.be	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 Techiques 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 β-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 β-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 β-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 χ 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 Moult, 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 [Plemper 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 Moult, 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, hydropathy, 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. AGGRESCAN: 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, Lehväslaiho 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 Aβ42 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, Csizmók 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, Csizmók 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, Sandor 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, Coggill 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 π-stacking in the self-assembly of amyloid fibrils. FASEB J 16:77–83.

Giardine B, Riemer C, Hefferon T, Thomas D, Hsu F, Zielenski J, Sang Y, Elnitski L, Cutting G, Trumbower H, Kern A, Kuhn R, Patrinos GP, Hughes J, Higgs D, Chui D, Scriver C, Phommarinh 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, Easteal S, Jermiin LS. 2007. Mind the gaps: evidence of bias in estimates of multiple sequence alignments. Mol Biol Evol 24:2433–2442.

Gorodkin J, Stærfeldt HH, Lund O, Brunak S. 1999. MatrixPlot: visualizing sequence constraints. Bioinformatics 15:769–770.

Gsponer J, Haberthur U, Caflisch 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.

Guerois 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.

Hyytinen 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 β2-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 Cα geometry: φ, ψ and Cβ 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.forcasp.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.

Plemper 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 α-complementing domain of *E. coli* β-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, Boronenkov IV, Anderson RA, Hurley JH. 1998. Structure of type IIâ 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. SNPeffect 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, Matouschek 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, Sorokine 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, Abeysinghe 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, Makhatadze 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–W672.

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, Caflisch 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.

Tompa P. 2002. Intrinsically unstructured proteins. Trends Biochem Sci 27: 527–533.

Tompa 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 Aβ (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.