Dear Author

Here are the proofs of your article.

- You can submit your corrections **online,** via **e-mail** or by **fax**.
- For **online** submission please insert your corrections in the online correction form. Always indicate the line number to which the correction refers.
- You can also insert your corrections in the proof PDF and **email** the annotated PDF.
- For **fax** submission, please ensure that your corrections are clearly legible. Use a fine black pen and write the correction in the margin, not too close to the edge of the page.
- Remember to note the **journal title**, **article number**, and **your name** when sending your response via e-mail or fax.
- **Check** the metadata sheet to make sure that the header information, especially author names and the corresponding affiliations are correctly shown.
- **Check** the questions that may have arisen during copy editing and insert your answers/corrections.
- **Check** that the text is complete and that all figures, tables and their legends are included. Also check the accuracy of special characters, equations, and electronic supplementary material if applicable. If necessary refer to the *Edited manuscript*.
- The publication of inaccurate data such as dosages and units can have serious consequences. Please take particular care that all such details are correct.
- Please **do not** make changes that involve only matters of style. We have generally introduced forms that follow the journal's style.
- Substantial changes in content, e.g., new results, corrected values, title and authorship are not allowed without the approval of the responsible editor. In such a case, please contact the Editorial Office and return his/her consent together with the proof.
- If we do not receive your corrections **within 48 hours**, we will send you a reminder.
- Your article will be published **Online First** approximately one week after receipt of your corrected proofs. This is the **official first publication** citable with the DOI. **Further changes are, therefore, not possible.**
- The **printed version** will follow in a forthcoming issue.

Please note

After online publication, subscribers (personal/institutional) to this journal will have access to the complete article via the DOI using the URL:

`http://dx.doi.org/10.1007/s12265-011-9259-1`

If you would like to know when your article has been published online, take advantage of our free alert service. For registration and further information, go to: http://www.springerlink.com.

Due to the electronic nature of the procedure, the manuscript and the original figures will only be returned to you on special request. When you return your corrections, please inform us, if you would like to have these documents returned.

Metadata of the article that will be visualized in OnlineFirst

Please note: Image will appear in color online but will be printed in black and white.

1	Article Title	**Meet Me Halfway: When Genomics Meets Structural Bioinformatics**	
2	Article Sub- Title		
3	Article Copyright - Year	**Springer Science+Business Media, LLC 2011 (This will be the copyright line in the final PDF)**	
4	Journal Name	Journal of Cardiovascular Translational Research	
5	Corresponding Author	Family Name	**Gong**
6		Particle	
7		Given Name	**Sungsam**
8		Suffix	
9		Organization	Cardiovascular Biomedical Research Unit, Royal Brompton Hospital
10		Division	
11		Address	Sydney Street, London SW3 6NP, UK
12		Organization	University of Cambridge
13		Division	Biocomputing Group, Department of Biochemistry
14		Address	80 Tennis Court Road, Cambridge CB2 1GA, UK
15		e-mail	s.gong@rbht.nhs.uk
16	Author	Family Name	**Worth**
17		Particle	
18		Given Name	**Catherine L.**
19		Suffix	
20		Organization	University of Cambridge
21		Division	Biocomputing Group, Department of Biochemistry
22		Address	80 Tennis Court Road, Cambridge CB2 1GA, UK
23		Organization	Charité Universitätsmedizin Berlin
24		Division	
25		Address	Lindenberger Weg 80, Berlin 13125, Germany
26		e-mail	
27	Author	Family Name	**Cheng**
28		Particle	
29		Given Name	**Tammy M. K.**
30		Suffix	
31		Organization	University of Cambridge
32		Division	Biocomputing Group, Department of Biochemistry
33		Address	80 Tennis Court Road, Cambridge CB2 1GA, UK

	Organization	Cancer Research UK London Research Institute
	Division	
	Address	44 Lincoln’s Inn Fields, London WC2A 3LY, UK
	e-mail	
Author	Family Name	**Blundell**
	Particle	
	Given Name	**Tom L.**
	Suffix	
	Organization	University of Cambridge
	Division	Biocomputing Group, Department of Biochemistry
	Address	80 Tennis Court Road, Cambridge CB2 1GA, UK
	e-mail	
Schedule	Received	30 November 2010
	Revised	
	Accepted	8 February 2011
Abstract		The DNA sequencing technology developed by Frederick Sanger in the 1970s established genomics as the basis of comparative genetics. The recent invention of next-generation sequencing (NGS) platform has added a new dimension to genome research by generating ultra-fast and high-throughput sequencing data in an unprecedented manner. The advent of NGS technology also provides the opportunity to study genetic diseases where sequence variants or mutations are sought to establish a causal relationship with disease phenotypes. However, it is not a trivial task to seek genetic variants responsible for genetic diseases and even harder for complex diseases such as diabetes and cancers. In such polygenic diseases, multiple genes and alleles, which can exist in healthy individuals, come together to contribute to common disease phenotypes in a complex manner. Hence, it is desirable to have an approach that integrates *omics* data with both knowledge of protein structure and function and an understanding of networks/pathways, i.e. functional genomics and systems biology; in this way, genotype–phenotype relationships can be better understood. In this review, we bring this ‘bottom-up’ approach alongside the current NGS-driven genetic study of genetic variations and disease aetiology. We describe experimental and computational techniques for assessing genetic variants and their deleterious effects on protein structure and function.
Keywords separated by ' - '		Next-generation sequencing - Genotype–phenotype relationship - Single nucleotide polymorphism - Computational algorithm - Database
Foot note information		

J. of Cardiovasc. Trans. Res.
DOI 10.1007/s12265-011-9259-1

Meet Me Halfway: When Genomics Meets Structural Bioinformatics

Sungsam Gong · Catherine L. Worth · Tammy M. K. Cheng · Tom L. Blundell

Received: 30 November 2010 / Accepted: 8 February 2011

Abstract The DNA sequencing technology developed by Frederick Sanger in the 1970s established genomics as the basis of comparative genetics. The recent invention of next-generation sequencing (NGS) platform has added a new dimension to genome research by generating ultra-fast and high-throughput sequencing data in an unprecedented manner. The advent of NGS technology also provides the opportunity to study genetic diseases where sequence variants or mutations are sought to establish a causal relationship with disease phenotypes. However, it is not a trivial task to seek genetic variants responsible for genetic diseases and even harder for complex diseases such as diabetes and cancers. In such polygenic diseases, multiple genes and alleles, which can exist in healthy individuals, come together to contribute to common disease phenotypes in a complex manner. Hence, it is desirable to have an approach that integrates *omics* data with both knowledge of protein structure and function and an understanding of networks/pathways, i.e. functional genomics and systems biology; in this way, genotype–phenotype relationships can be better understood. In this review, we bring this 'bottom-up' approach alongside the current NGS-driven genetic study of genetic variations and disease aetiology. We describe experimental and computational techniques for assessing genetic variants and their deleterious effects on protein structure and function.

Keywords Next-generation sequencing · Genotype–phenotype relationship · Single nucleotide polymorphism · Computational algorithm · Database

S. Gong · C. L. Worth · T. M. K. Cheng · T. L. Blundell
Biocomputing Group, Department of Biochemistry,
University of Cambridge,
80 Tennis Court Road,
Cambridge CB2 1GA, UK

S. Gong (✉)
Cardiovascular Biomedical Research Unit,
Royal Brompton Hospital,
Sydney Street,
London SW3 6NP, UK
e-mail: s.gong@rbht.nhs.uk

C. L. Worth
Charité Universitätsmedizin Berlin,
Lindenberger Weg 80,
13125 Berlin, Germany

T. M. K. Cheng
Cancer Research UK London Research Institute,
44 Lincoln's Inn Fields,
London WC2A 3LY, UK

Introduction

The year 2010 marked the tenth anniversary of the completion of the first draft of the entire human genome. During this time, there has been remarkable progress in genome biology, mainly achieved by next-generation sequencing (NGS) technologies of unprecedented throughput [1, 2], by which the human genome is assembled and annotated in ways unthinkable a decade ago. The challenge now is to try to understand the genetic differences, many of which presently remain unknown, which underlie healthy and unhealthy human individuals. It has become evident that for many Mendelian diseases, single nucleotide polymorphisms (SNPs) connect phenotype with genotype and are potentially valuable for understanding the mechanism of diseases. In the future, we need to understand SNPs that contribute to more complex polygenic diseases. Many of these occur in healthy individuals in the population and therefore are likely to have less impact on the structures and functions of individual gene products. Most are in non-coding regions between genes that control other genes, but usually, their functions are not known. A few are in protein coding regions, but many of these appear to be to be 'neutral'. Only one or two may affect function—are 'drivers'—but how can they be identified? The huge numbers of SNPs make the challenge greater; so far, there are about 6.5 million SNPs in public databases.

Although the pace of collecting SNP data is impressive, progress in annotating SNPs is relatively slow. For example, the National Center for Biotechnology Information (NCBI) dbSNP database [3], which is a major repository of human SNPs, contained data for about 4.8 million human SNPs in total as of Build 106 in 2002. By Build 132 in October 2010, there were about 143 million human SNPs. However, there are only a few thousand SNPs that have been associated with a human genetic disorder in the Online Mendelian Inheritance in Man (OMIM) database [4, 5]. The identification of disease-associated SNPs via informatics approaches has become a major challenge which requires urgent attention.

The Association Between Genetic Variation and Disease

A SNP is technically defined as a sequence variant that occurs at <99% frequency in the population [6]. It represents the most widespread type of sequence variation in the human genome and is estimated to be found at a rate of 1 in every 200–300 bp [7, 8] and 1 in every 1,000 bp in coding regions of genes [9]. Extrapolation to the entire genome suggests that at least 15 million SNPs exist, about three orders of magnitude greater than the number of genes. They probably account for 90% of genetic variation in man [10], and more importantly, they affect individual susceptibility to diseases and responses to drugs. For example, several disease studies, such as Alzheimer's disease, support the idea that individuals with specific SNPs are more likely to suffer from a disease compared to the rest of the population [11]. Also, genetic predisposition has been associated with several cases of drug hypersensitivity [12].

However, for the reasons given above, understanding the contribution of SNPs to disease remains challenging, and the results are often controversial. There are two extreme hypotheses, both relying on the idea that common genetic variants underlie susceptibility to common diseases [13]. One of them suggests that a comprehensive collection of SNPs, especially those within coding and regulatory regions, will include common variants that can influence disease susceptibility (SNP–disease association) [14], whereas the other emphasises the allelic association with nearby SNPs which can be used as genetic markers for a certain disease (allele–disease association) [15]. From the SNP–disease association point of view, studying the relation of SNPs to protein structure is a good approach for learning about protein structure and function, and it has been shown useful for structure-based drug design [16]. From the allelic–disease association point of view, linkage disequilibrium (LD) concerns the association of alleles in a specific population compared to a random distribution. In the case of identifying disease-associated alleles, the comparison of marker allele frequencies between patients and healthy individuals via LD analysis has been useful for narrowing the region of the genome where the disease gene must lie [17]. Because LD studies are able to identify lower risk disease-associated variants, they may be powerful in analysing complex diseases as well [14, 18].

In terms of genetic effects, SNPs can cause various effects according to their location in the genome. They may cause gene regulation abnormities when they appear in transcription initiation sites, or may affect mRNA splicing when they appear at the splice site between intron donor and acceptors (so-called regulatory SNPs). A small proportion of SNPs will cause an amino acid change in the protein sequence, known as non-synonymous SNPs (nsSNPs). These nsSNPs may directly affect the normal function of proteins by altering binding sites in proteins such as protein, nucleic acid, ligand or ion binding sites. Protein function may also be affected by nsSNPs that alter protein stability, protein aggregation or posttranslational modifications (leading to improper trafficking of the protein). In either case, where protein malfunction occurs, disease may result. Thus, nsSNPs are important sources for understanding the mechanism of diseases: nearly half of the human monogenic Mendelian diseases may be accounted for by nsSNPs [19]. The challenge is to obtain a similar understanding for more polygenic, non-Mendelian diseases.

Amino Acid Variations and Diseases

Most monogenic diseases, such as sickle cell disease and severe combined immunodeficiency, appear to result from a single DNA variant resulting in an amino acid substitution, which affects protein stability rather than impairing protein function directly [20, 21]. Therefore, methods that predict the effects of mutations on protein stability are useful for identifying possible disease associations [22, 23]. Indeed, several computer programmes are designed to identify protein mutations that affect protein stability, which we cover later in this article. However, for most common diseases such as cancers, heart diseases and diabetes, where multiple genes and alleles play a role in complex phenotypes or traits, pinpointing the genetic loci underlying continues to be challenging. With the help of recent advances in sequencing technologies [24, 25] and analytical frameworks (see [26, 27] for review), we are now beginning to see successful case studies identifying the genetic loci underlying the aetiology of complex diseases such as type 1 [28, 29] and 2 diabetes [30, 31], asthma and coronary heart disease [32, 33]. More recently, systematic resequencing of the cancer genome has revealed genetic changes that may be responsible for lung, breast and colorectal cancers [34–37]. Lists of genetic loci associated with disease susceptibility from published studies are deposited in databases such as T1Dbase [38], COSMIC [39], the European Genome–Phenome Archive) [40], ModSNP [41], SwissVar [42], HGMD [19] and a Catalog of Published Genome-Wide Association Studies[1] of the National Human Genome Research Institute (NHGRI). Table 1 lists public database resources compiled for human genetic variations, their effects on protein structures and disease information. Therefore, our understanding of the genetic basis of diseases is beginning to improve with the help of large-scale genome-wide association studies (GWAS) and high-throughput sequencing technologies, although more molecular and physiological studies of genetic variants need to follow in order to confirm association with disease aetiology.

Insights Gained from Mendelian Diseases

Before the determination of the human genome sequence, analysis of genetic mutations focused on establishing the relationship between genotype and phenotype, especially susceptibility to certain disease types [43, 44]. However, there were no general methods for identifying DNA sequences responsible for even simple Mendelian diseases until 1980 when Botstein and colleagues developed a method for constructing a linkage map of the human genome, with restriction fragment length polymorphisms as molecular markers [45, 46]. After this initial milestone, the human genetic linkage map and the methods and algorithms have been used for connecting disease genes, traits or mutations with Mendelian diseases and have been successful in the identification of 1,200 disease genes, including the classic examples of sickle cell anaemia [47], hemochromatosis [48] and lactose intolerance [49] (see [17, 50] for reviews).

Detailed molecular analyses of protein structure and function have revealed that single amino acid substitutions or mutations are often responsible for certain disease types [20, 51]. It has been claimed that ~60% of such Mendelian disease mutations arise from amino acid substitutions in their respective genes (see [17] for a review). For most monogenic diseases, a single DNA variant resulting in an amino acid substitution is responsible for a certain disease type by affecting protein stability and, thus, function [21]. Table 2 shows four types of disease variants classified by the functional effect on their final gene products with examples of Mendelian diseases. Hence, much effort has been expended to characterise the pattern of mutations in the context of sequences and structures of proteins in attempts to establish whether they are likely to be neutral or deleterious to the functions of the organism [52–55]. Interestingly, most of the methods that aim to assess deleterious mutations are based on principles observed in nature—to see whether the mutations conform to the neutral theory of protein evolution [56–59] which selects against radical changes of amino acids. However, real challenges at present are from complex diseases that obscure the genetic basis responsible for molecular phenotypes.

Challenges from Complex Diseases

Linkage mapping, as mentioned earlier, has been successful in identifying the genetic basis of Mendelian diseases, such as Huntington disease [60] and cystic fibrosis [61, 62] where the relationship between genotype and phenotype is straightforward and the diagnosis is unequivocal due to the monogenic nature of the diseases [17]. However, even before the first linkage map was completed, it was recognized that most human traits and diseases follow complex modes of inheritance. Hence, it is not a trivial task to identify the genetic traits or variants responsible for complex diseases such as cancers and diabetes where the phenotypes are determined by the combination of multiple genes and interactions between genes and environmental factors that can affect gene expression.

These difficulties have started to be resolved by technical advancements in modern sequencing methods (see [24] for a review) which enable the charting of genetic variation between human individuals in a fast and highly accurate manner. A seminal project initiated from the Wellcome Trust Case Control Study[2] harnesses the power of such genotyping technologies to improve our understanding of the aetiological basis of complex diseases such as type 1 diabetes, type 2 diabetes, coronary heart disease, hypertension, bipolar disorder, rheumatoid arthritis and Crohn’s disease. For each disease type, genome sequence variations, SNPs in particular, are mapped by comparing the genetic makeup of the case group (disease) and the control group (normal). This allows the identification of many SNPs and genes that have evidence of association with disease susceptibility [26, 33, 50, 63]. In addition, the ENCODE project[3] (ENCyclopedia Of DNA Elements)

[1] http://www.genome.gov/gwastudies/.

[2] http://www.wtccc.org.uk/.

[3] http://www.genome.gov/10005107.

Table 1 List of databases compiled for human genetic variations and diseases

Name	URL	Summary	Reference
HGMD	http://www.hgmd.cf.ac.uk/ac/index.php	A comprehensive core collection of data on published germline mutations in nuclear genes underlying human inherited disease	[19]
dbSNP	http://www.ncbi.nlm.nih.gov/projects/SNP/	A free public archive for genetic variation within and across different species developed and hosted by the NCBI in collaboration with the NHGRI	[3]
HGVbase (HGBASE)	http://www.hgvbaseg2p.org/	A catalogue of all known sequence variations (particularly SNPs) as a non-redundant set of records, which presents each variant in the context of its physical relationship to the nearest human gene	[193, 194]
ProTherm	http://gibk26.bse.kyutech.ac.jp/jouhou/Protherm/protherm.html	A collection of numerical data of thermodynamic parameters such as Gibbs free energy change, enthalpy change, heat capacity change, transition temperature etc. for wild-type and mutant proteins, which are important for understanding the structure and stability of proteins	[195]
ASEdb	www.asedb.org	A repository for energetics of side chain interactions determined by alanine-scanning mutagenesis	[196]
p53	http://www.bioinf.org.uk/p53/	Integrating mutation data and structural analysis of p53 tumour-suppressor protein	[197]
G6PD	http://www.bioinf.org.uk/g6pd/	An integration of up-to-date mutational and structural data of human glucose-6-phosphate dehydrogenase (G6PD) from various genetic and structural databases (Genbank, Protein Data Bank, etc.) and latest publications	[198]
MutDB	http://mutdb.org/	Annotation of human variation data with protein structural information and other functionally relevant information	[199]
SNPper	http://snpper.chip.org/	A web-based application designed to facilitate the retrieval and use of human SNPs for high-throughput research purposes	[200]
ModSNP (SwissVar)	http://expasy.org/swissvar/	A portal to search variants in Swiss-Prot entries of the UniProt Knowledgebase (UniProtKB) and gives direct access to the Swiss-Prot Variant pages	[41, 42]
COSMIC	http://www.sanger.ac.uk/genetics/CGP/cosmic/	To store and display somatic mutation information and related details and contains information relating to human cancers	[39, 201]
TopoSNP	http://gila.bioengr.uic.edu/snp/toposnp/	An interactive visualisation of disease and non-disease associated nsSNPs and displays geometric and relative entropy calculations	[202]
LS-SNP	http://salilab.org/LS-SNP/	A genomic-scale, computational pipeline that maps human SNPs in NCBI's dbSNP database [3] onto protein sequences in the SwissProt/TrEMBL databases	[203]
SAAPdb	http://www.bioinf.org.uk/saap/	Integration of information on single amino acid polymorphisms (i.e. structurally expressed SNPs and mutations) with analysis of the likely structural effects of these amino acid mutations	[204]
SNPeffect	http://snpeffect.vib.be/	Annotations for both non-coding and coding SNP, as well as annotations for the SwissProt set of human disease mutations.	[205, 206]
SNP@Domain	http://snpnavigator.net	A web resource of single nucleotide polymorphisms (SNPs) within protein domain structures and sequences	[207]
T1DBase	http://t1dbase.org	A public web site and database that supports the type 1 diabetes (T1D) research community.	[38]
PolyDoms	http://polydoms.cchmc.org/polydoms/	A database to integrate the results of multiple algorithmic procedures and functional criteria applied to the entire Entrez dbSNP dataset. In addition to predicting structural and functional impacts of all nsSNPs, filtering functions enable group-based identification of potentially harmful nsSNPs among multiple genes associated with specific diseases, anatomies, mammalian phenotypes, gene ontologies, pathways or protein domains	[208]
DMDM	http://bioinf.umbc.edu/dmdm/	A database in which each disease mutation can be displayed by its gene, protein or domain location. DMDM provides a unique domain-level view where all human coding mutations are mapped on the protein domain	[209]
DVGa	http://www.ebi.ac.uk/dgva/page.php	A public catalogue of the large-scale insertions, deletions, duplications and rearrangements that are found in the genomes of individuals within a species	[40]
1,000 Genome	http://www.1000genomes.org	The project aims to find most genetic variants that have frequencies of at least 1% in the populations studied by sequencing many individuals lightly	[64]
WTCCC	http://www.wtccc.org.uk/	To exploit progress in understanding of patterns of human genome sequence variation along with advances in high-throughput genotyping technologies, and to explore the utility, design and analyses of GWA studies	[210]

Table 2 Classification of mutation types by effect on functions

Types of variants	Definitions	Example			
		Disease	OMIM	Gene	Descriptions
Loss-of-function	A mutation that produces a reduced amount/activity of product or a complete loss of function (an amorphic mutation)	Phenylketonuria (PKU)	261600	*PAH*	Mutations often destabilise hepatic enzymephenylalaninehydroxylase (PAH) which is responsible for phenylalanine metabolism
Gain-of-function	A mutation that produces a increased amount/activity or novel product which may do something positively abnormal (a hypermorph or neomorph)	Polyostotic fibrous dysplasia (PFD)	174800	*GNAS1*	Mutation at the gene induce stimulatory G-protein alpha subunit (G_s-α), a key component of many signal transduction pathways
Dominant negative	A mutation which makes its product antagonizes (interferes) the activity of the normal product (an antimorph)	Romano–Ward syndrome (RWS) or cardiac arrhythmia (Long QT syncrome)	192500	*KCNQ1*	The dominantly inherited Romano-Ward syndrome is caused by dominant negative mutations in the voltage-gated potassium channel-1 gene
Haploinsufficiency	A specific case of loss-of-function mutation where the reduced dosage of a normal gene product (with the other copy inactivated by loss-of-function) is not enough for a normal phenotype in a diploid organism	Waardenburg syndrome type 1 (WS1)	193500	*PAX3*	An autosomal dominant auditory-pigmentary syndrome characterised by pigmentary abnormalities of the hair, skin, and eyes

aims to identify all functional elements in the human genome sequence, and the 1,000 Genome Project[4] aims to construct the most accurate human genetic variation map to support disease studies [64]. These international efforts look very promising, but there is still a long way to go to establish a complete understanding of disease mechanisms, especially at the molecular level.

Experimental Techniques for Investigating Effect(s) of Mutations on Proteins

Site-directed mutagenesis allows a mutation to be introduced at a specific position in a DNA molecule, for example, a plasmid. The effect (if any) that the mutation has on protein stability or function can then be verified by comparison to the wild-type protein; for an overview of early work in this area, see the following excellent reviews [65, 66].

The oligomeric state of wild-type and mutant proteins can be determined by carrying out a range of experiments such as chemical cross-linking, size exclusion chromatography, analytical ultracentrifugation, dynamic light scattering, gel filtration chromatography and ion mobility–mass spectrometry analysis [67–69]. The secondary and tertiary structure content of proteins can be determined using far-UV CD measurements. Protein unfolding can be induced using chemical denaturants (e.g. urea or guanidinium chloride) or by increasing temperature and the changes monitored using fluorescence and far-UV CD measurements [70–72]. Comparison of the wild-type and mutant spectra will reveal whether there are differences in stability. The reversibility of unfolding can also be investigated using fluorescence and far-UV CD measurements where folding is reversible and the renaturation curve will superimpose on the denaturation curve. The folding kinetics of proteins can be studied using stopped-flow, single-jump or double-jump experiments monitored by fluorescence or far-UV CD [73, 74]. Where a mutant protein has essentially the same tertiary structure as the wild-type protein, the local effects of the mutation can be assessed using hydrogen–deuterium exchange experiments [75].

The functional effects of mutations can be investigated by measuring the effect of a mutation on binding to other proteins, e.g. co-immunoprecipitation followed by Western blotting [76, 77], or to ligands [78]. The rate at which enzyme catalysis occurs can be measured, for example, by pre-steady-state inhibition kinetics to determine whether a mutation is having an effect [79, 80]. Where functional assays are available, these can be used to determine the effect of mutation on protein function.

Determining the experimental structure of wild-type and/or mutant proteins can provide further insight into the mechanism underlying observed biophysical or functional

[4] http://www.1000genomes.org.

differences [81–85]. In short, a detailed biophysical and functional analysis of a protein can be achieved by carrying out a combination of these techniques, and where disease-associated mutations are known, these techniques can help identify the mechanism underlying the disease, particularly where the structure of the protein is known [86, 87].

The tumour suppressor protein, p53, is an example of a protein that has been extensively characterised using many of the techniques outlined above. Bullock et al. [88] were able to carry out reversible denaturation of the p53 core domain and mutants using urea denaturation at 10°C, showing that five oncogenic mutations cause various degrees of destabilization depending on whether the residues are associated with DNA binding, zinc ion binding or structural contacts. Structural mutations that thermodynamically destabilise p53 were also shown to be kinetically destabilising, resulting in faster unfolding rates in comparison to wild-type protein [89].

Protein engineering of p53 was achieved using naturally occurring amino acid substitutions in p53 homologues from 23 species along with known second-site suppressor mutations, resulting in a superstable quadruple mutant that has almost the same binding affinity for the gadd45 promoter as the wild-type protein [90]. Later work showed that these four mutations cause only small structural changes, with two of the mutations increasing the rigidity of the structure and none affecting the overall structure of the DNA-binding surface or β-sandwich [91]. The superstable quadruple mutant was further used to identify the molecular mechanism underlying two oncogenic mutants, R273H and R249S, and the latter mutant's rescue by oncogenic mutant H168R [92]. The crystal structures of these oncogenic mutants revealed that R273H abolished an arginine involved with DNA binding, whereas H168R and R249S induce large structural changes in loop regions [92]. The authors further demonstrated that the arginine introduced by H168R mimics the role of that lost in R249S, restoring the wild-type conformation of the loop regions. It was later demonstrated that mutant R273H weakens specific DNA binding by ~700–1,000 times compared to that of wild-type p53 [93].

These types of experiments are of course limited to investigating the functional effects of mutations in vitro. In order to understand the physiological effect of mutations, they need to be tested in vivo, for instance by using genetically engineered animal models [94–96].

Computational Methods to Assess Genetic Mutations

The types of detailed experiments described above are invaluable for characterising the structural and/or functional effects of mutations in proteins. However, the sheer volume of SNP data that have been generated in recent years from projects such as the Human Genome Project [97], HapMap Project [98] and genome-wide association studies means that it is not possible to characterise all nsSNPs in such a way. Automatic methods that can predict the effect that a mutation will have on a protein are required; accurate predictions will allow a reduced set of SNPs to be characterised experimentally, saving time and money. In the next few paragraphs and in Table 3, we describe software that is currently available for testing the effect of nsSNPs in silico; however, this is by no means a comprehensive survey, and the following review articles will provide further details [99–101].

It has been estimated that up to 80% of disease-associated nsSNPs are caused by protein stabilization effects [20], although this analysis was carried out on monogenic disease and therefore the same pattern may not be observed with complex diseases. Nonetheless, methods that predict the effect that nsSNPs will have on protein stability are useful for identifying possible disease associations.

The structural effects of nsSNPs can be assessed using methods for predicting stability changes in proteins using sequence and/or structural information: site-directed mutator (SDM) uses observed substitutions in homologous protein families in the form of probability tables to calculate a stability score [102]—we revisit this method in detail later in this article; FoldEF (also known as FoldX) uses weighted empirical data in the form of energy terms to calculate stability differences between wild-type and mutant proteins [103]; I-MUTANT is a neural network method that can be used to predict whether a mutation is stabilising or destabilising [104]. The same authors also produced a support vector machine (SVM) tool, I-MUTANT2.0, for predicting protein stability changes caused by mutation that can use either sequence or structural information [105, 106]. A similar approach is employed by MUpro where three SVMs have been developed based on sequence only, structure only and sequence and structural information [107]. CUPSAT uses structural environment-specific atom potentials and torsion angles to predict stability changes between wild-type and mutant proteins [108], and ERIS uses physical descriptions of atomic interactions and models backbone flexibility to predict stability changes upon mutation [109].

Some of the programmes detailed above may be able to distinguish mutations that lead to protein aggregation from neutral mutations. However, various methodologies are now available for specifically predicting protein aggregation: TANGO is a statistical mechanics algorithm based on the physicochemical principles of β-sheet formation [110]; AGGRESCAN is a web-based programme that identifies putative aggregation hot spots based on the aggregation propensity of each residue in a protein sequence [111]; PASTA uses a pairwise energy function based on the

Table 3 Computer software and web applications to study the effects genetic mutations and disease associations

Name	URL	Summary	Reference
SDM	http://www-cryst.bioc.cam.ac.uk/~sdm/sdm.php	A statistical potential energy function developed to predict the effect that mutations on the stability of proteins	[23, 55, 211]
PopMuSiC	http://babylone.ulb.ac.be/popmusic/	A statistical potential approach for the computer-aided design of mutant proteins with controlled stability properties. It evaluates the changes in stability of a given protein or peptide under single-site mutations, on the basis of the protein's structure	[212, 213]
SIFT	http://sift.jcvi.org/	An algorithm taking a query sequence and using multiple alignment information to predict tolerant and deleterious substitutions for every position of the query sequence	[53, 121]
DFIRE	http://sparks.informatics.iupui.edu/yueyang/DFIRE/dDFIRE-service	Distance-scaled, finite ideal gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction	[214]
FOLDEF	http://foldx.crg.es/	A computer algorithm to provide a fast and quantitative estimation of the importance of the interactions contributing to the stability of proteins and protein complexes using an empirical potential approach	[103, 215]
Polyphen	http://genetics.bwh.harvard.edu/pph/	A tool which predicts possible impact of an amino acid substitution on the structure and function of a human protein using straightforward physical and comparative considerations	[140, 216]
I-mutant	http://gpcr2.biocomp.unibo.it/cgi/predictors/I-Mutant/I-Mutant.cgi	A neural network method that can be used to predict whether a mutation is stabilising or destabilising	[104]
Panther	http://www.pantherdb.org/tools/	A library of protein families and subfamilies derived by the use of hidden Markov model techniques indexed by a vocabulary of more than 500 biological functional terms (aka. subPSEC)	[139]
GROMOS	http://www.igc.ethz.ch/GROMOS/index	A force field for molecular dynamics simulation.	[217]
I-mutant 2.0	http://gpcr2.biocomp.unibo.it/cgi/predictors/I-Mutant2.0/I-Mutant2.0.cgi	SVN (Support Vector Machine) version of I-mutant.	[106]
PHD-SNP	http://gpcr.biocomp.unibo.it/~emidio/PhD-SNP/PhD-SNP.htm	A decision tree with the SVM-based classifier coupled to the SVM-Profile trained on sequence profile information	[125]
nsSNPAnalyzer	http://snpanalyzer.utmem.edu/	Web-based software which extracts structural and evolutionary information from a query nsSNP and uses a machine learning method called Random Forest to predict the nsSNP's phenotypic effect (the web is down at the time of this writing)	[123]
Pmut	http://mmb2.pcb.ub.es:8080/PMut/	Computer software aimed at the annotation and prediction of pathological mutations by retrieving a series of structural parameters such as volume parameters, secondary structure propensities, hydrophobicity descriptors and sequence potential, among others	[218]
Mupro	http://mupro.proteomics.ics.uci.edu/	A machine-learning approach based on support vector machines (SVMs) to predict the stability changes for single site mutations	[107]
CUPSAT	http://cupsat.tu-bs.de/	A tool to predict changes in protein stability upon point mutations	[108]
FastSNP	http://fastsnp.ibms.sinica.edu.tw/pages/input_CandidateGeneSearch.jsp	An web-based application which prioritizes SNPs according to 12 phenotypic risks and putative functional effects, such as changes to the transcriptional level, pre-mRNA splicing, protein structure, etc.	[219]
SNPs3D	http://www.snps3d.org/	A web site which assigns molecular functional effects of non-synonymous SNPs based on structure and sequence analysis	[220]
ERIS	http://troll.med.unc.edu/eris/login.php	The Eris server calculates the change of the protein stability induced by mutations ($\Delta\Delta G$) utilizing the recently developed Medusa modelling suite	[221]
SAPRED	http://sapred.cbi.pku.edu.cn/	An automatic pipeline to predict the disease association of SAPs using several novel attributes such as Structural Neighbor Profile and Nearby Functional Sites, in addition to incorporating other well-known attributes such as Residue Frequency and Conservation	[222]
stSNP	http://ilyinlab.org/StSNP/	The structure SNP (StSNP) web server compares structural nsSNP distributions in many proteins or protein complexes. StSNP enables researchers to map nsSNPs onto protein structures by comparative modelling of structure with nsSNPs by MODELLER (http://salilab.org)	[223]

t3.23 **Table 3** (continued)

Name	URL	Summary	Reference
		and visualise their structural locations by using the multiple structure-sequence viewer Friend. Pathway information is provided from KEGG database	
t3.23 SNAP	http://snap.humgen.au.dk/views/index.cgi	A sequence analysis web server providing a simple but detailed analysis of human genes and their variations	[224]
t3.24 AUTO-MUTE	http://proteins.gmu.edu/automute/	A combined approach to predict stability changes in protein mutants based on a four-body, knowledge-based and statistical contact potential, and machine-learning techniques	[225]
t3.25 Bongo	www.bongo.cl.cam.ac.uk/Bongo/	A graph theoretic measure for estimation of structural and pathological impacts of non-synonymous SNP	[138]
t3.26 Omidios (SeqProfCod)	http://sgu.bioinfo.cipf.es/services/Omidios/	The Omidios web site takes a query SWISS-PROT id and searches for all annotated and predicted protein variants (nsSNP)	[226]
t3.27 F-SNP	http://compbio.cs.queensu.ca/F-SNP/	It provides integrated information about the functional effects of SNPs obtained from 16 bioinformatics tools and databases	[227]
t3.28 CHARMM	http://www.charmm.org/	A force field for molecular dynamics as well as the name for the molecular dynamics simulation and analysis package associated with them	[228]

propensities of two residues that face one another on neighbouring strands of a β-sheet to calculate energy scores to specific β-pairings of two sequence stretches of the same length [112].

The impact of nsSNPs on protein function can be assessed by identifying those residues in a protein structure that are involved in binding. The evolutionary trace (ET) method predicts active sites and functional interfaces in proteins using sequence conservation patterns in homologous proteins to extract functionally important residues. These residues can then be mapped on to the protein surface to generate spatial clusters identifying functional sites. Various servers exist which are based on this approach, including ET VIEWER [113], JEVTRACE [114], MINER [115] and CRESCENDO [116], the latter being able to distinguish between residues that are conserved due to structural restraints from those due to functional restraints. Active site residues can also be identified using databases of structural templates compiled from known catalytic residues in experimentally determined structures, for example, the Catalytic Site Atlas [117], PDBSiteScan [118] and MSDsite [119]. However, none of these functional site prediction programmes assess the impact that the mutant residue will have on function. The effect that a mutation will have on protein–protein binding energy is considered by FoldEF [103] and can be calculated using the molecular modelling programme Rosetta [120].

There are also numerous resources available for predicting whether a nsSNP will be deleterious or not: SIFT calculates a scaled probability for a substitution from an alignment of homologous proteins [53, 121]; POLYPHEN uses functional annotations, structural parameters and evolutionary information to predict whether a mutation will be deleterious [122]; nsSNPANALYZER uses a similar approach incorporating evolutionary information from sequence alignments with information from the three-dimensional structure of the protein of interest to predict the effect of the mutation [123]; SNAP is a neural network-based method that uses sequence information as input but includes functional and structural annotations [124]; PHD-SNP is based on a decision tree with the SVM-based classifier coupled to the SVM-Profile trained on sequence profile information [125].

In the following sections, we will focus on two computer programmes—SDM and Bongo—and a database, SAMUL, developed by our research group to exemplify how the use of protein structural data can help better understand genetic mutations underlying disease phenotypes.

SDM: A Statistical Potential Energy Function to Predict the Effect of Mutations on Protein Stability

In its native state, a polypeptide chain is folded in a specific 3D structure usually involving regions of secondary structure and specific tertiary interactions. However, under conditions of stress, e.g. high temperature, the protein may denature to an unfolded state, which is more flexible and highly hydrated. The thermodynamic stability of a protein is the difference in free energy between the folded and unfolded states and is thus a measure of the ratio of the two forms of the protein. Protein stability differences between wild-type and mutant proteins can be calculated using the thermodynamic cycle (see Fig. 1).

SDM[5] is a statistical potential energy function developed by Topham et al. [102] to predict the effects that mutations will have on the stability of proteins. SDM uses environment-specific amino acid substitution frequencies within homolo-

[5] http://www-cryst.bioc.cam.ac.uk/~sdm/sdm.php.

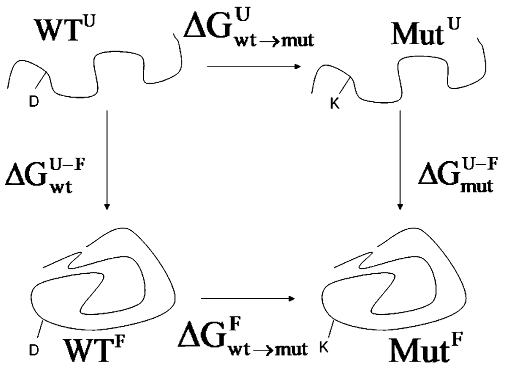


Fig. 1 The thermodynamic cycle can be used to calculate protein stability changes between wild-type and mutant proteins. The difference in free energy of unfolding of the wild type (*wt*) and mutant (*mut*), ΔΔG, can be calculated using this cycle. ΔG^{U-F}_{mut} and ΔG^{U-F}_{wt} represent the free energy change going from the unfolded (*U*) to the folded (*F*) state for the mutant and wild-type proteins, respectively. Direct simulation of the unfolding is not possible. As the total free energy in the full cycle is zero, the $\Delta\Delta G$ can instead be calculated using the free energy changes associated with the transformation of wt→mut in the unfolded and folded state ($\Delta G^U_{wt \to mut}$ and $\Delta G^F_{wt \to mut}$, respectively)

gous protein families [126, 127] to calculate a stability score which is analogous to the free energy difference between a wild-type and mutant protein. Blind testing on a set of 83 staphylococcal nuclease and 63 barnase mutants showed a correlation of 0.80 between the predicted stability changes and experimental data [102]. The method performs comparably or better than other published methods in the task of classifying mutations as stabilising or destabilising [55]. Additionally, SDM has a much improved sensitivity in predicting stabilising mutations compared to other published methods (five of the seven methods incorrectly classify >68% of the stabilising mutations). Therefore, SDM is a useful tool for helping to design site-directed mutagenesis experiments or for predicting whether a mutation will impact protein structure and have a role in disease. It has been employed in the analysis of mutations in the autoimmune regulator protein [128], mixed lineage kinase 3 [129], the adaptor protein MyD88 adaptor-like [130] and breast cancer susceptibility gene 1 [131]. Figure 2 displays an example of the output produced by the SDM server.

SDM has also been applied to the task of predicting deleterious nsSNPs at the genome scale [22, 55]. By combining the predictions made by SDM with functional site predictions by CRESCENDO [132] and observed interaction sites stored in the databases PICCOLO, BIPA [133] and CREDO [134], the structural and functional effects of nsSNPs can be differentiated, thereby potentially aiding the identification of the causative mechanism of a disease [23]. For instance, the software and database tools were used to generate new hypotheses regarding (1) the molecular aetiology of renal cell carcinoma and pheochromocytoma in the cancer syndrome, von Hippel–Lindau disease [135]; (2) the structural effects of mutations in thyroid-stimulating hormone receptor that are associated with congenital non-goitrous hypothyroidism [136]; and (3) tumour risk associated with mutations in succinate dehydrogenase D [137].

Bongo: An Application of Graph Theory to Predict Deleterious Mutations

Bonds ON Graph[6] (Bongo) [138] (Fig. 3) is a graph-based method that analyses the likelihood of a point mutation to cause diseases by affecting its corresponding protein structures. It considers a target protein as a residue–residue interaction graph in which vertices represent residues and edges represent interactions between residues, and applies graph theoretic measures to estimate the topological change due to single point mutations.

Identifying Structurally Important Key Residues in Proteins

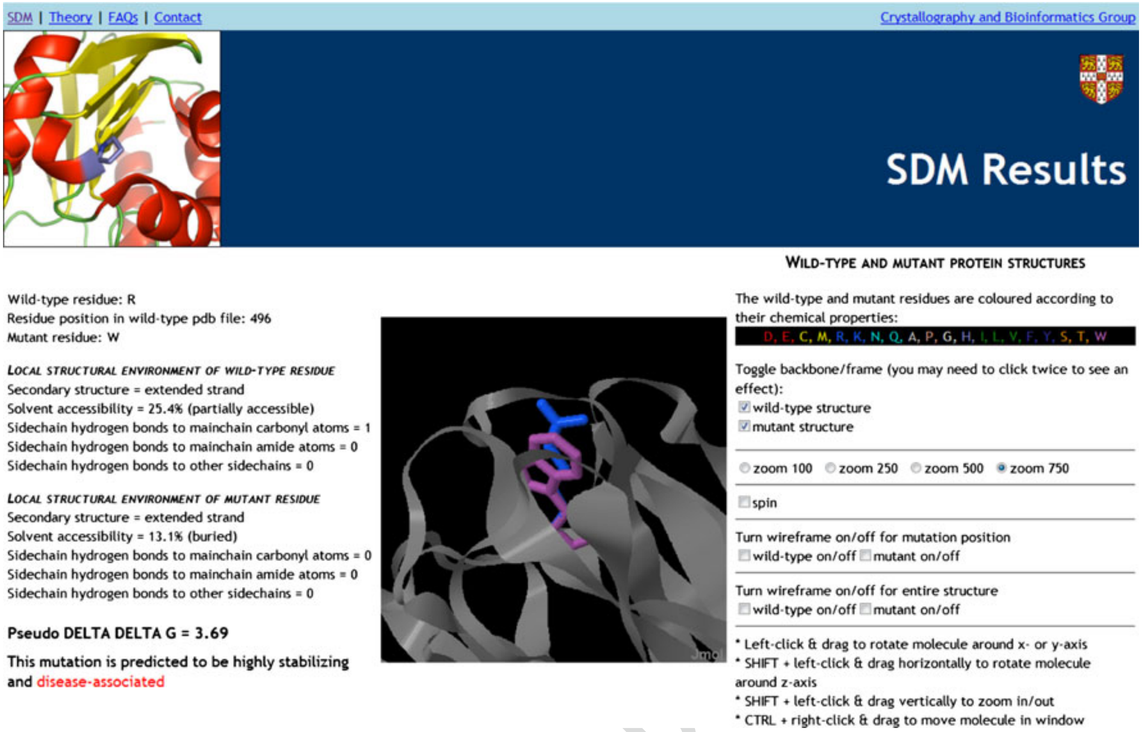
Bongo applies the idea of *vertex cover*, defined in graph theory as a minimum set of vertices (residues) that are crucial to forming all the edges (interactions), to identify key residues that play important roles in stabilising folded protein structures. The algorithm Bongo uses to select key residues is similar to the concept of pulling out one piece each time in a tower of wooden pieces (in this case the interconnected residue network), with the difference that in this case, the pieces pulled out are key pieces but not redundant ones and those pulled out at a earlier stage are supposed to be more significant in maintaining the overall network (see [138] for more details).

For a target mutation, Bongo identifies two sets of key residues from the residue interaction network of its corresponding wild-type and mutant protein structure. Then, Bongo quantifies the structural effect of the mutation via comparing the difference of the two key residue sets.

Estimating Stability Change by Investigating Protein Structure Alone

Changing key residues in the protein internal network reflects stability changes contributed by amino acid variations, as shown in previous studies of point mutations in the p53 core domain, barnase, human and bacteriophage T4 lysozyme, gene V protein of bacteriophage f1 and ribonuclease H from *Escherichia coli* [138]. This echoes the underlying concept of Bongo, formulated on the basis that disease-causing mutations often affect intrinsic structural features of proteins [51] and most disease-associated mutations appear to affect protein stability rather than

[6] http://www.bongo.cl.cam.ac.uk/Bongo/.



SDM | Theory | FAQs | Contact
Crystallography and Bioinformatics Group
SDM Results

Wild-type and mutant protein structures

Wild-type residue: R
Residue position in wild-type pdb file: 496
Mutant residue: W

Local structural environment of wild-type residue
Secondary structure = extended strand
Solvent accessibility = 25.4% (partially accessible)
Sidechain hydrogen bonds to mainchain carbonyl atoms = 1
Sidechain hydrogen bonds to mainchain amide atoms = 0
Sidechain hydrogen bonds to other sidechains = 0

Local structural environment of mutant residue
Secondary structure = extended strand
Solvent accessibility = 13.1% (buried)
Sidechain hydrogen bonds to mainchain carbonyl atoms = 0
Sidechain hydrogen bonds to mainchain amide atoms = 0
Sidechain hydrogen bonds to other sidechains = 0

Pseudo DELTA DELTA G = 3.69

This mutation is predicted to be highly stabilizing and disease-associated

The wild-type and mutant residues are coloured according to their chemical properties:
D, E, C, M, R, K, N, Q, A, P, G, H, I, L, V, F, Y, S, T, W

Toggle backbone/frame (you may need to click twice to see an effect):
☑ wild-type structure
☑ mutant structure

○ zoom 100 ○ zoom 250 ○ zoom 500 ● zoom 750

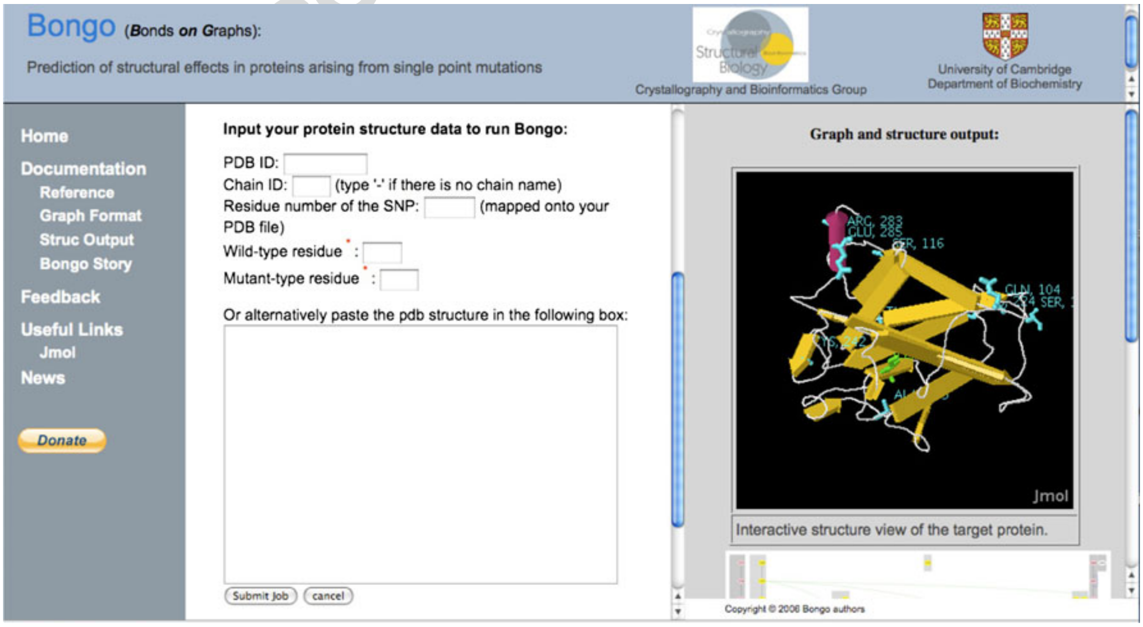
□ spin

Turn wireframe on/off for mutation position
□ wild-type on/off □ mutant on/off

Turn wireframe on/off for entire structure
□ wild-type on/off □ mutant on/off

- Left-click & drag to rotate molecule around x- or y-axis
- SHIFT + left-click & drag horizontally to rotate molecule around z-axis
- SHIFT + left-click & drag vertically to zoom in/out
- CTRL + right-click & drag to move molecule in window

Fig. 2 Screen shot of SDM. The SDM results for mutation R496W in proprotein convertase subtilisin/kexin type 9 (PDB code 2P4E) are shown. On the *left-hand side*, information about the local structural environment of the wild-type and mutant residues is provided, such as secondary structure and hydrogen bonds. Below this information, the stability score prediction is shown as well as a prediction of whether the mutation is disease-associated or not. In this instance, the mutation is predicted to be highly stabilising and disease-associated. In fact, this mutation has been found in high/low-density lipoprotein cholesterol subjects [229] and has been implicated in hypercholesterolemia [230]. In the *middle*, the Jmol viewer is shown, displaying the wild-type and mutant residues. On the *right-hand side*, various options are provided for manipulating how the two structures are displayed, such as zooming in and viewing only one of the two structures



Prediction of structural effects in proteins arising from single point mutations

Input your protein structure data to run Bongo:

PDB ID:
Chain ID: (type '-' if there is no chain name)
Residue number of the SNP: (mapped onto your PDB file)
Wild-type residue * :
Mutant-type residue * :

Or alternatively paste the pdb structure in the following box:

Graph and structure output:

Interactive structure view of the target protein.

Fig. 3 Screen dump of Bongo. Bongo is currently available as a web server for public use. Users can either copy and paste their structures or use a PDB code as an input

interfere directly with protein interactions [20]. This fundamental idea was also further validated on a benchmark set which contains 506 disease-associated nsSNPs from the OMIM database [4] and 220 non-disease-associated nsSNPs available in the dbSNP database [3], over which Bongo has positive predictive value and negative predictive value of 78.5% and 34.5%, respectively, comparable to that of PolyPhen 1 [122] and PANTHER [139].

Another specific feature of Bongo is its pure prerequisite of structural information—it considers protein internal network alone, and thus, the prediction result is complimentary to other contemporary approaches, such as SDM [102], SIFT [53, 121] and PolyPhen 2 [140], that apply protein sequence or a combination of sequence and structure information. At the same time, this makes it possible to annotate further the mechanism of point mutations in proteins: we can estimate whether a point mutation has greater impact on function or structure (or equally on both parts) by cross-referencing the analytical results of methods that apply only sequence or structure information.

SAMUL: A Database for Systematic Annotations of Macromolecules

SAMUL[7] provides comprehensive structural and functional annotations of amino acid residues and accommodates amino acid variations and mutations so that they can be browsed and interpreted in conjunction with the structural and functional environments of the wild-type amino acid residues. Through the mapping between sequence and structures, information about protein–protein interaction (PICCOLO) [141], protein–nucleic acid interaction (BIPA) [133], protein–ligand interaction (CREDO) [134] and nsSNPs on protein structure are interconnected [55, 142]. The structural annotations are mainly from the local structural environments of amino acid residues determined by JOY [143] and presented and highlighted using Jmol, a molecular viewer [144]. For functional annotations, UniProt feature descriptions are selected and transferred onto their corresponding positions in 3D structures if available.

Rich Annotations on Amino Acid Variations

Thirty-four annotations are provided at amino acid residue level, of which six are structural annotations of 3D structures and the remaining 28 are functional annotations mainly from UniProt feature (FT) descriptions. Table 4 shows the full list of annotations available from SAMUL. SAMUL houses amino acid sequence variants from *Homo sapiens* genome annotation provided by the following data sources: (1) Ensembl human variation database [145], (2) cancer somatic mutations from COSMIC [39] and (3) UniProt human sequence variations [146]. They are integrated with various annotation information mentioned in the previous section. Table 5 shows the number of SNPs mapped onto UniProt, Protein Data Bank (PDB), PICCOLO, CREDO and BIPA at the time of writing. SNPs in Ensembl proteins were mapped onto their corresponding UniProt proteins and further to proteins in the PDB via Double-map [147]. Among them, nsSNPs are of special interest, especially if their allele types change corresponding amino acids, which are presumably responsible for interactions in PICCOLO, CREDO and BIPA.

Visualisation of Annotations

GBrowse Structural and functional annotations are graphically displayed and highlighted at the residue level of UniProt (or Ensembl) protein sequence using Generic Genome Browser (GBrowse), which is an open-source genome viewer widely used in the community [148]. Figure 4 shows a GBrowse-generated image highlighting functional and structural annotations of a cell division protein kinase 2 (CDK2, UniProt accession no. P24941). The image can be locally saved in various formats such as PNG, SVG and PDF through the web site. Annotations on the image are linked to the original sources of information so that users can investigate those features in depth.

Jmol Structural and functional annotations mapped onto the 3D structure of PDB files can be selected and highlighted within the Jmol macromolecular viewer [144]. Figure 5 shows a Jmol-embedded SAMUL screenshot displaying the 3D structure of cell division protein kinase 2 (CDK2, PDB code no. 1E1V), highlighting the location of various structural and functional features.

Distributed Annotation System SAMUL is a Distributed Annotation System (DAS) server which provides XML-based web services to disseminate structural and functional annotations through the web. The DAS protocol is built on a client–server system which allows a single machine to communicate distant web server to gather up different types of biological annotations, collate the information and display it to the end user in a single view. Most of the major knowledge-based biological systems such as Ensembl, UCSC genome browser [149] and WormBase [150] provide DAS services. Numerous DAS resources are coordinated by the DAS registration server[8] [151].

[7] http://www-cryst.bioc.cam.ac.uk/samul (or http://samul.org, alternatively).

[8] http://www.dasregistry.org/.

Table 4 List of structural and functional annotations provided from SAMUL (TLB for the in-house resource developed in the TLB group)

Source	Annotations	URL	Descriptions
The Blundell Group[a]	PICCOLO	http://www-cryst.bioc.cam.ac.uk/piccolo	Protein–protein interaction database
	CREDO	http://www-cryst.bioc.cam.ac.uk/credo	A protein–ligand interaction database for drug discovery
	BIPA	http://www-cryst.bioc.cam.ac.uk/bipa	Biological interaction database for protein–nucleic acid
UNIPROT	REGION	http://www.uniprot.org/manual/region	Extent of a region of interest in the sequence
	VAR_SEQ	http://www.uniprot.org/manual/var_seq	Description of sequence variants produced by alternative splicing, alternative promoter usage, alternative initiation and ribosomal frameshifting
	VARIANT	http://www.uniprot.org/manual/variant	Authors report that sequence variants exist
	HUMSAVAR	http://www.uniprot.org/docs/humsavar	Human polymorphisms and disease mutations
	TRANSMEM	http://www.uniprot.org/manual/transmem	Extent of a transmembrane region
	NP_BIND	http://www.uniprot.org/manual/np_bind	Extent of a nucleotide phosphate-binding region
	MUTAGEN	http://www.uniprot.org/manual/mutagen	Site which has been experimentally altered by mutagenesis
	DISULFID	http://www.uniprot.org/manual/disulfid	Cysteine residues participating in disulfide bonds
	METAL	http://www.uniprot.org/manual/metal	Binding site for a metal ion
	DNA_BIND	http://www.uniprot.org/manual/dna_bind	Denotes the position and type of a DNA-binding domain
	MODRES	http://www.uniprot.org/manual/mod_res	Modified residues excluding lipids, glycans and protein cross-links
	BINDING	http://www.uniprot.org/manual/binding	Binding site for any chemical group (co-enzyme, prosthetic group, etc.)
	ZN_FING	http://www.uniprot.org/manual/zn_fing	Denotes the position(s) and type(s) of zinc fingers within the protein
	ACT_SITE	http://www.uniprot.org/manual/act_site	Amino acid(s) directly involved in the activity of an enzyme
	PEPTIDE	http://www.uniprot.org/manual/peptide	Extent of an active peptide in the mature protein
	MOTIF	http://www.uniprot.org/manual/motif	Short (up to 20 amino acids) sequence motif of biological interest
	COMPBIAS	http://www.uniprot.org/manual/compbias	Region of compositional bias in the protein
	CARBOHYD	http://www.uniprot.org/manual/carbohyd	Covalently attached glycan group(s)
	CA_BIND	http://www.uniprot.org/manual/ca_bind	Denotes the position(s) of calcium binding region(s) within the protein
	PROPEP	http://www.uniprot.org/manual/propep	Part of a protein that is cleaved during maturation or activation
	SITE	http://www.uniprot.org/manual/site	Any interesting single amino acid site on the sequence
	SIGNAL	http://www.uniprot.org/manual/signal	Sequence targeting proteins to the secretory pathway or periplasmic space
	TRANSIT	http://www.uniprot.org/manual/transit	Extent of a transit peptide for organelle targeting
	CROSSLNK	http://www.uniprot.org/manual/crosslnk	Residues participating in covalent linkage(s) between proteins
	NON_TER	http://www.uniprot.org/manual/non_ter	The sequence is incomplete. Indicate that a residue is not the terminal residue of the complete protein
	LIPID	http://www.uniprot.org/manual/lipid	Covalently attached lipid group(s)
CSA	CSA_PSI	http://www.ebi.ac.uk/thornton-srv/databases/CSA/	A database documenting enzyme active sites and catalytic residues in enzymes of 3D structure: homologous entries, found by PSI-BLAST alignment to one of the original entries
	CSA_LIT	http://www.ebi.ac.uk/thornton-srv/databases/CSA/	A database documenting enzyme active sites and catalytic residues in enzymes of 3D structure: original hand-annotated entries, derived from the primary literature
COSMIC	COSMIC	http://www.sanger.ac.uk/genetics/CGP/cosmic/	Catalogue of Somatic Mutations In Cancer

Table 4 (continued)

Source	Annotations	URL	Descriptions
ENSEMBL	ENVAR	http://www.ensembl.org/info/docs/variation/index.html	Ensembl Human variation database
PDB	MOD_RES	http://www.wwpdb.org/documentation/format32/sect3.html#MODRES	descriptions of modifications (e.g., chemical or posttranslational) to protein and nucleic acid residues

[a] http://www-cryst.bioc.cam.ac.uk/

Bridging the Gap Between Genome and Proteome Through Structure Modelling

Since the first determination of a protein sequence, that of insulin by Sanger and Tuppy in the 1950s [152, 153], high-throughput sequencing techniques have enabled massive production of sequence information from different organisms. UniProt [154] is a central hub for protein sequences, providing rich annotation on function and cross-references. In parallel, recent technical advancements in X-ray crystallography and NMR experiments have enabled massive production of protein structure information. The PDB is the main repository of 3D structures of biological protein macromolecules [155]. By knowing the structure of a protein, we are able to gain a better understanding of its function, and this in turn allows the development of pharmacological agents to manipulate its activity. Despite the ongoing structural genomics projects that aim to rectify the deficit of protein structures, there is still a massive discrepancy in the number of protein sequences in genomic databases compared to protein structures [156], mainly because transmembrane proteins are underrepresented in the PDB (see Fig. 6). Hence, not all protein coding variants can be analysed in the context of 3D structure. However, even where a structure is not available for a particular protein, a reliable model structure can be built if a homologous protein (with sequence identity of 30% or more) has an experimentally determined structure [157, 158], which can help increase the coverage of nsSNPs to be analysed within the structural context. This process is called comparative modelling, and over the past 20 years, it has become well established as a reliable means of generating a hypothetical structure of a protein in the absence of experimental structures [159].

Table 5 Number of distinct SNPs categorized by annotations in SAMUL

Type	Database	No. of distinct SNPs
Sequence	Ensembl	203,484
	UniProt	194,053
Structure	PDB	18,963
	PICCOLO	4,696
	CREDO	3,263
	TOPO_DOM	3,068
	REGION	2,412
	ZN_FING	183
	NP_BIND	140
	DNA_BIND	135
	BIPA	122
	PEPTIDE	115
	COSMIC	110
	DISULFID	100
	MOD_RES	92
	CSA_PSI	85
	CARBOHYD	81
	MUTAGEN	71
	SITE	63
	BINDING	62
	COMPBIAS	53
	MODRES	52
	TRANSMEM	47
	METAL	45
	PROPEP	42
	CA_BIND	37
	MOTIF	37
	ACT_SITE	23
	CROSSLNK	5
	CSA_LIT	4
	NON_TER	3
	TRANSIT	2
	SIGNAL	1

Comparative modelling requires a template sequence(s) that is structurally homologous to the protein of interest. As mentioned previously, tertiary structure tends to be more conserved than primary structure; therefore, even proteins with relatively low sequence identity to the query protein can be used as template structures if they are structurally homologous. Template proteins can be identified using sequence alignment methods [160, 161], although it will often be necessary to employ methods that can identify more distant relationships using, for example, profile hidden Markov models [162], threading methods [163] and environment-specific substitution table-based methods [164]. A range of programmes exists for building a model structure based on the sequence alignment, and approaches are broadly classed as either restraint-based [165–167] or

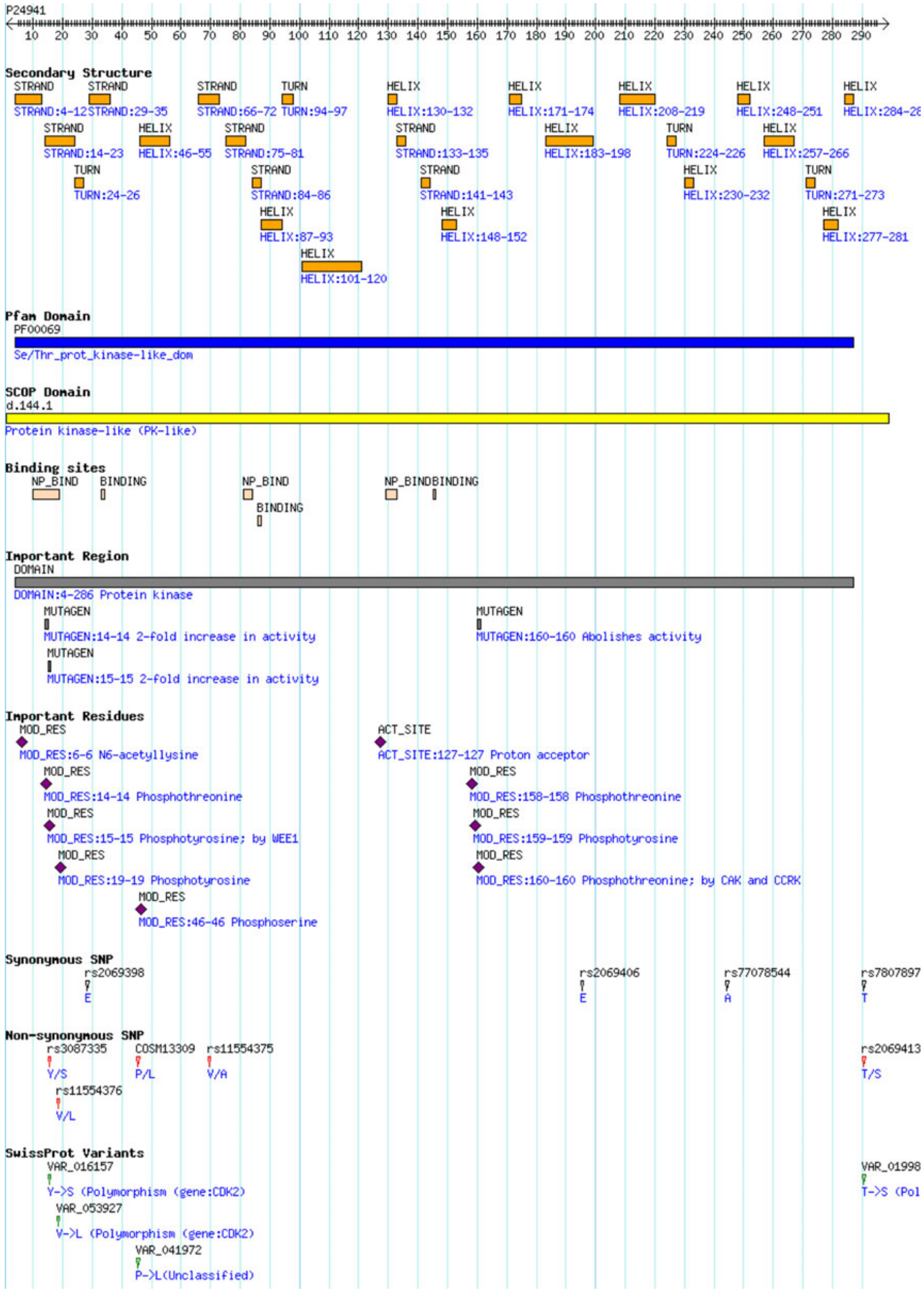
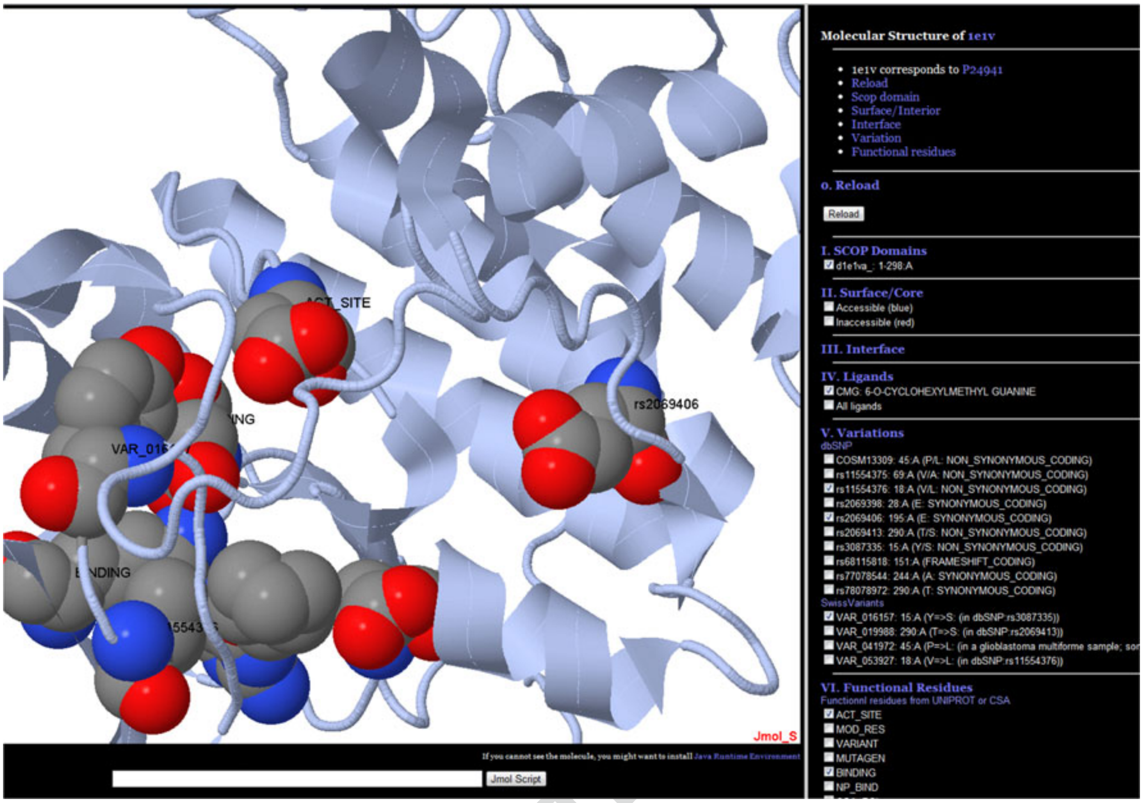

P24941
10 20 30 40 50 60 70 80 90 100 110 120 130 140 150 160 170 180 190 200 210 220 230 240 250 260 270 280 290
Secondary Structure
STRAND STRAND STRAND TURN HELIX HELIX HELIX HELIX HELIX
STRAND:4-12 STRAND:29-35 STRAND:66-72 TURN:94-97 HELIX:130-132 HELIX:171-174 HELIX:208-219 HELIX:248-251 HELIX:284-28
STRAND HELIX STRAND STRAND HELIX TURN HELIX
STRAND:14-23 HELIX:46-55 STRAND:75-81 STRAND:133-135 HELIX:183-198 TURN:224-226 HELIX:257-266
TURN STRAND STRAND HELIX TURN
TURN:24-26 STRAND:84-86 STRAND:141-143 HELIX:230-232 TURN:271-273
HELIX HELIX HELIX
HELIX:87-93 HELIX:148-152 HELIX:277-281
HELIX
HELIX:101-120
Pfam Domain
PF00069
Se/Thr_prot_kinase-like_dom
SCOP Domain
d.144.1
Protein kinase-like (PK-like)
Binding sites
NP_BIND BINDING NP_BIND NP_BINDBINDING
BINDING
Important Region
DOMAIN
DOMAIN:4-286 Protein kinase
MUTAGEN
MUTAGEN:14-14 2-fold increase in activity
MUTAGEN
MUTAGEN:160-160 Abolishes activity
MUTAGEN
MUTAGEN:15-15 2-fold increase in activity
Important Residues
MOD_RES
MOD_RES:6-6 N6-acetyllysine
ACT_SITE
ACT_SITE:127-127 Proton acceptor
MOD_RES
MOD_RES:14-14 Phosphothreonine
MOD_RES
MOD_RES:158-158 Phosphothreonine
MOD_RES
MOD_RES:15-15 Phosphotyrosine; by WEE1
MOD_RES
MOD_RES:159-159 Phosphotyrosine
MOD_RES
MOD_RES:19-19 Phosphotyrosine
MOD_RES
MOD_RES:160-160 Phosphothreonine; by CAK and CCRK
MOD_RES
MOD_RES:46-46 Phosphoserine
Synonymous SNP
rs2069398 rs2069406 rs77078544 rs7807897
E E A T
Non-synonymous SNP
rs3087335 COSM13309 rs11554375 rs2069413
Y/S P/L V/A T/S
rs11554376
V/L
SwissProt Variants
VAR_016157
Y->S (Polymorphism (gene:CDK2)
VAR_01998
T->S (Pol
VAR_053927
V->L (Polymorphism (gene:CDK2)
VAR_041972
P->L(Unclassified)

Fig. 4 A screen shot (http://www-cryst.bioc.cam.ac.uk/gb2/gbrowse/samul/?name=P24941) of GBrowse from SAMUL Structural and functional annotations of human cell division protein kinase 2 (CDK2, UniProt: P24941) are shown along with the protein sequence shown at the top of the figure. The annotations consist of nine tracks: (1) secondary structure, (2) Pfam [162] and (3) SCOP [231] for domain assignment information, (4) binding sites, (5) important regions and (6) important sites for functional features, (7) synSNP, (8) nsSNP and (9) SwissVariants [146] for amino acid variation information. The functional features are from UniProt [154] and amino acid variations are from dbSNP [3]



Molecular Structure of 1e1v

- 1e1v corresponds to P24941
- Reload
- Scop domain
- Surface/Interior
- Interface
- Variation
- Functional residues

o. Reload

I. SCOP Domains
☑ d1e1va_: 1-298:A

II. Surface/Core
□ Accessible (blue)
□ Inaccessible (red)

III. Interface

IV. Ligands
☑ CMG: 6-O-CYCLOHEXYLMETHYL GUANINE
□ All ligands

V. Variations
dbSNP
□ COSM13309: 45:A (P/L: NON_SYNONYMOUS_CODING)
□ rs11554375: 69:A (V/A: NON_SYNONYMOUS_CODING)
☑ rs11554376: 18:A (V/L: NON_SYNONYMOUS_CODING)
□ rs2069398: 28:A (E: SYNONYMOUS_CODING)
☑ rs2069406: 195:A (E: SYNONYMOUS_CODING)
□ rs2069413: 290:A (T/S: NON_SYNONYMOUS_CODING)
□ rs3087335: 15:A (Y/S: NON_SYNONYMOUS_CODING)
□ rs68115818: 151:A (FRAMESHIFT_CODING)
□ rs77078544: 244:A (A: SYNONYMOUS_CODING)
□ rs78078972: 290:A (T: SYNONYMOUS_CODING)
SwissVariants
☑ VAR_016157: 15:A (Y=>S: (in dbSNP:rs3087335))
□ VAR_019988: 290:A (T=>S: (in dbSNP:rs2069413))
□ VAR_041972: 45:A (P=>L: (in a glioblastoma multiforme sample; sor
□ VAR_053927: 18:A (V=>L: (in dbSNP:rs11554376))

VI. Functional Residues
Functionnl residues from UNIPROT or CSA
☑ ACT_SITE
□ MOD_RES
□ VARIANT
□ MUTAGEN
☑ BINDING
□ NP_BIND

If you cannot see the molecule, you might want to install Java Runtime Environment

Fig. 5 Screen shot(http://www-cryst.bioc.cam.ac.uk/samul/pdb/1E1V/jmol?bgcolor=white) of Jmol from SAMUL. Structural and functional annotations of a human CDK2 (PDB: 1E1V) are shown in the 3D structure of the protein. The navigation panel is on the *right-hand side* and the Jmol viewer is on the *left*. The predefined structural and functional annotations are presented as follows: (1) SCOP domain, (2) surface and core regions, (3) interface residues (if any) between two adjacent SCOP domain, (4) types of ligand, (5) amino acid variants and (6) functional residues from the UniProt entry. There is also a form input field which accepts Jmol queries for advanced users who wish to manipulate visualisation options with their own flavours. The main chain of the protein molecule is presented as a *cartoon with structural annotations in space-filled models of the individual amino acids*

fragment-based [168–171]. Comparative models can be validated by checking for disallowed conformations in the structure [172], comparing the observed amino acid substitution patterns in the alignment with those predicted from environment-specific substitution tables [164] or by using statistical potentials to assess residue–residue contact frequencies [173].

Recent advances have also been made in high-resolution ab initio protein structure prediction where a model structure is built using ‘first principles’ and without the use of a template structure [174]. However, the successes reported tend to be limited to small proteins (<100 residues) [159], probably in part due to the vast amounts of computing power required to successfully carry out these simulations.

Limitations and Challenges

In this article, we reviewed recent advancements in genetic studies behind disease phenotypes and revisited various techniques for interrogating genetic variants responsible for disease aetiology. We mainly focused on genetic variants located in protein coding regions and those that result in amino acid substitutions. However, complex diseases are not always influenced by coding SNPs. Indeed, more evidence is emerging for the role of intronic SNPs that control splicing and expression (and timing of expression) of DNA and RNA products [175, 176], and even synonymous SNPs are reported to control mRNA stability and for correct splicing [177]. Indeed, based on Ensembl human variation database version 57, reported SNPs comprise 0.46% (0.13% for verified SNPs) of the total number of human DNA base pairs, of which 53% of SNPs occur at intergenic regions and 36% occur at intronic region (see Table 6). Only 1.26% of human SNPs occur in protein coding regions in which more than half are non-synonymous SNPs (0.64%) and the rest are synonymous SNPs (0.46%), frameshift (0.09%) and stop gained mutations (0.02%). We did not discuss those genetic variants responsible for insertions and deletions of DNA bases [178, 179] and larger copy number variants [180–182] because

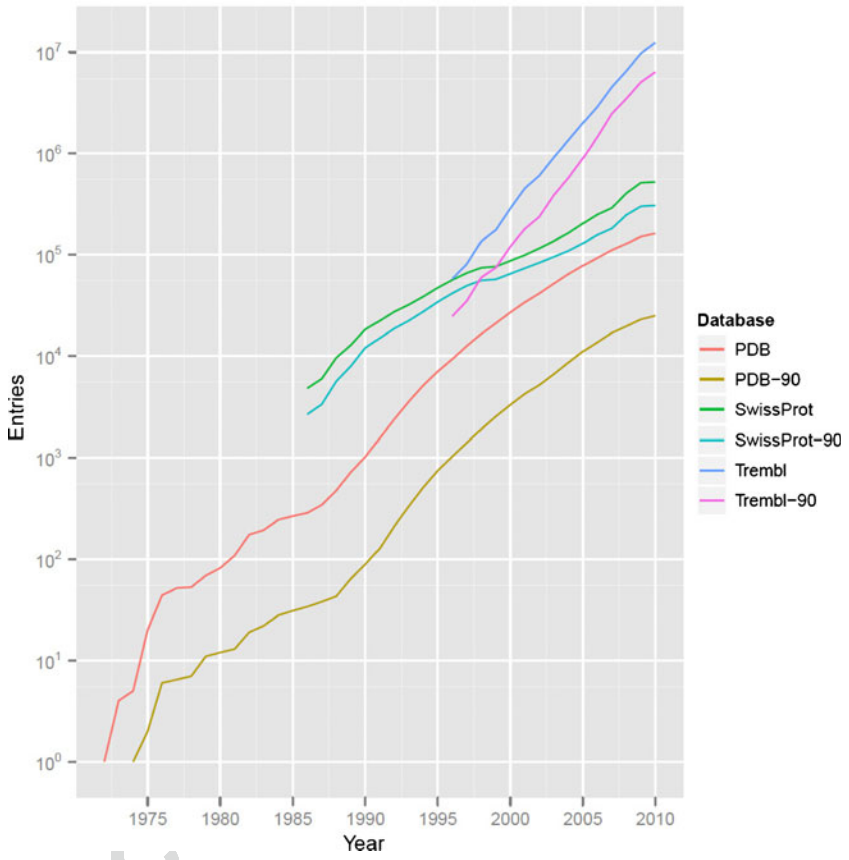


Fig. 6 Growth of biological databases. The number of entries deposited in SwissProt, TrEMBL and PDB are plotted by year. The SwissProt entries are manually annotated and curated, whereas those of TrEMBL are unreviewed protein sequences associated with computationally generated annotation from large-scale genome sequences. The PDB is a repository of proteins whose 3D structures are determined mainly by X-ray crystallography or NMR. Note that the number of PDB entry is counted based on protein chains because one PDB entry can have multiple protein sequences (e.g. oligomers or multimers). Entries in SwissProt, TrEMBL are PDB are clustered by their sequence identity of 90% or more, and they are shown as SwisProt-90, TrEMBL-90 and PDB-90, respectively

Table 6 Total number of SNPs by different types of their consequences

Type	Occurrence	Ratio (%)
INTERGENIC	7,982,768	53.07
INTRONIC	5,481,863	36.45
UPSTREAM	663,985	4.41
DOWNSTREAM	556,742	3.70
3PRIME_UTR	137,639	0.92
NON_SYNONYMOUS_CODING	96,031	0.64
WITHIN_NON_CODING_GENE	86,955	0.58
SYNONYMOUS_CODING	69,035	0.46
5PRIME_UTR	28,343	0.19
FRAMESHIFT_CODING	14,002	0.09
REGULATORY_REGION,INTRONIC	13,365	0.09
SPLICE_SITE,INTRONIC	10,457	0.07
REGULATORY_REGION,UPSTREAM	4,951	0.03
REGULATORY_REGION,INTERGENIC	4,949	0.03
NON_SYNONYMOUS_CODING, SPLICE_SITE	2,845	0.02
STOP_GAINED	2,533	0.02

Data from Ensemble human variations version 57

they are more difficult to study in terms of protein 3D structure; these have been reviewed elsewhere [183–188].

In addition, we also did not take the expression level into account; rather, we assumed that proteins are expressed equally no matter whether they contain sequence variants or not. However, it is clear that proteins having deleterious mutations are selectively controlled by the protein degradation system to protect against misfolded or damaged proteins [189], and sometimes those mutations are compensated in other species [190]. A recently developed tool for transcriptomics study can be harnessed to improve our understanding of genetic variations on the protein expression level [191, 192].

The predicted candidate variants underlying disease aetiology still need further molecular studies to verify the significance of in silico analysis and the structural bioinformatics approach. However, one major gain of this reductionist approach is that it could be complementary to current genetics-based statistical approaches such as GWAS by prioritizing genetic variants for further validation. Considering one of the critiques of GWAS, which states that it does not provide any functional implication of genetic variation [63], the structural analysis of amino acid variants could be advantageous and indeed applicable to molecular diagnosis and further to personalized medicine, particularly if genetics and structural bioinformatics meet halfway.

Acknowledgement We are grateful to all who are developing and maintaining biological databases, scientists submitting their invaluable data and people who support open-source programmes and operating systems. SG, CLW and TC want to thank the Blundell Group members for their support and collaborations during their stay in Cambridge.

References

1. Metzker, M. L. (2010). Sequencing technologies—The next generation. *Nature Reviews. Genetics, 11*(1), 31–46.
2. Wheeler, D. A., Srinivasan, M., Egholm, M., Shen, Y., Chen, L., McGuire, A., et al. (2008). The complete genome of an individual by massively parallel DNA sequencing. *Nature, 452* (7189), 872–876.
3. Sherry, S. T., Ward, M. H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., et al. (2001). dbSNP: The NCBI database of genetic variation. *Nucleic Acids Research, 29*(1), 308–311.
4. Hamosh, A., Scott, A. F., Amberger, J., Bocchini, C., Valle, D., & McKusick, V. A. (2002). Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Research, 30*(1), 52–55.
5. Hamosh, A., Scott, A. F., Amberger, J. S., Bocchini, C. A., & McKusick, V. A. (2005). Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Research, 33*, D514–517.
6. Schork, N. J., Fallin, D., & Lanchbury, J. S. (2000). Single nucleotide polymorphisms and the future of genetic epidemiology. *Clinical Genetics, 58*(4), 250–264.
7. Kruglyak, L., & Nickerson, D. A. (2001). Variation is the spice of life. *Nature Genetics, 27*(3), 234–236.
8. Stephens, J. C., Schneider, J. A., Tanguay, D. A., Choi, J., Acharya, T., Stanley, S. E., et al. (2001). Haplotype variation and linkage disequilibrium in 313 human genes. *Science, 293*(5529), 489–493.
9. Chakravarti, A. (1998). It's raining SNPs, hallelujah? *Nature Genetics, 19*(3), 216–217.
10. Collins, F. S., Brooks, L. D., & Chakravarti, A. (1998). A DNA polymorphism discovery resource for research on human genetic variation. *Genome Research, 8*(12), 1229–1231.
11. Emahazion, T., Feuk, L., Jobs, M., Sawyer, S. L., Fredman, D., St Clair, D., et al. (2001). SNP association studies in Alzheimer's disease highlight problems for complex disease analysis. *Trends in Genetics, 17*(7), 407–413.
12. Pirmohamed, M. (2006). Genetic factors in the predisposition to drug-induced hypersensitivity reactions. *The AAPS Journal, 8* (1), E20–26.
13. Lohmueller, K. E., Pearce, C. L., Pike, M., Lander, E. S., & Hirschhorn, J. N. (2003). Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. *Nature Genetics, 33*(2), 177–182.
14. Risch, N., & Merikangas, K. (1996). The future of genetic studies of complex human diseases. *Science, 273*(5281), 1516–1517.
15. Kruglyak, L. (1999). Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nature Genetics, 22*(2), 139–144.
16. Tsigelny, I. F., Kotlovyi, V., & Wasserman, L. (2004). SNP analysis combined with protein structure prediction defines structure–functional relationships in cancer related cytochrome P450 estrogen metabolism. *Current Medicinal Chemistry, 11*(5), 525–538.
17. Botstein, D., & Risch, N. (2003). Discovering genotypes underlying human phenotypes: past successes for Mendelian disease, future approaches for complex disease. *Nature Genetics, 33*(Suppl), 228–237.
18. Lander, E. S. (1996). The new genomics: Global views of biology. *Science, 274*(5287), 536–539.
19. Stenson, P. D., Ball, E. V., Mort, M., Phillips, A. D., Shiel, J. A., Thomas, N. S., et al. (2003). Human Gene Mutation Database (HGMD): 2003 update. *Human Mutation, 21*(6), 577–581.
20. Wang, Z., & Moult, J. (2001). SNPs, protein structure, and disease. *Human Mutation, 17*(4), 263–270.
21. Yue, P., Li, Z., & Moult, J. (2005). Loss of protein structure stability as a major causative factor in monogenic disease. *Journal of Molecular Biology, 353*(2), 459–473.
22. Burke, D. F., Worth, C. L., Priego, E. M., Cheng, T., Smink, L. J., Todd, J. A., et al. (2007). Genome bioinformatic analysis of nonsynonymous SNPs. *BMC Bioinformatics, 8*, 301.
23. Worth, C. L., Bickerton, G. R., Schreyer, A., Forman, J. R., Cheng, T. M., Lee, S., et al. (2007). A structural bioinformatics approach to the analysis of nonsynonymous single nucleotide polymorphisms (nsSNPs) and their relation to disease. *Journal of Bioinformatics and Computational Biology, 5*(6), 1297–1318.
24. Mardis, E. R. (2008). Next-generation DNA sequencing methods. *Annual Review of Genomics and Human Genetics, 9*, 387–402.
25. Morozova, O., Hirst, M., & Marra, M. A. (2009). Applications of new sequencing technologies for transcriptome analysis. *Annual Review of Genomics and Human Genetics, 10*, 135–151.
26. McCarthy, M. I., Abecasis, G. R., Cardon, L. R., Goldstein, D. B., Little, J., Ioannidis, J. P., et al. (2008). Genome-wide association studies for complex traits: Consensus, uncertainty and challenges. *Nature Reviews. Genetics, 9*(5), 356–369.
27. Weir, B. S. (2008). Linkage disequilibrium and association mapping. *Annual Review of Genomics and Human Genetics, 9*, 129–142.
28. Hakonarson, H., Grant, S. F., Bradfield, J. P., Marchand, L., Kim, C. E., Glessner, J. T., et al. (2007). A genome-wide association study identifies KIAA0350 as a type 1 diabetes gene. *Nature, 448*(7153), 591–594.
29. Todd, J. A., Walker, N. M., Cooper, J. D., Smyth, D. J., Downes, K., Plagnol, V., et al. (2007). Robust associations of four new chromosome regions from genome-wide analyses of type 1 diabetes. *Nature Genetics, 39*(7), 857–864.
30. Sladek, R., Rocheleau, G., Rung, J., Dina, C., Shen, L., Serre, D., et al. (2007). A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature, 445*(7130), 881–885.
31. Zeggini, E., Weedon, M. N., Lindgren, C. M., Frayling, T. M., Elliott, K. S., Lango, H., et al. (2007). Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. *Science, 316*(5829), 1336–1341.
32. Burton, P. R., Clayton, D. G., Cardon, L. R., Craddock, N., Deloukas, P., Duncanson, A., et al. (2007). Association scan of 14,500 nonsynonymous SNPs in four diseases identifies autoimmunity variants. *Nature Genetics, 39*(11), 1329–1337.
33. Consortium WTCC. (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature, 447*(7145), 661–678.
34. Dalgliesh, G. L., Furge, K., Greenman, C., Chen, L., Bignell, G., Butler, A., et al. (2010). Systematic sequencing of renal carcinoma reveals inactivation of histone modifying genes. *Nature, 463*(7279), 360–363.
35. Greenman, C., Stephens, P., Smith, R., Dalgliesh, G. L., Hunter, C., Bignell, G., et al. (2007). Patterns of somatic mutation in human cancer genomes. *Nature, 446*(7132), 153–158.
36. Pleasance, E. D., Cheetham, R. K., Stephens, P. J., McBride, D. J., Humphray, S. J., Greenman, C. D., et al. (2010). A comprehensive

catalogue of somatic mutations from a human cancer genome. *Nature, 463*(7278), 191–196.

37. Wood, L. D., Parsons, D. W., Jones, S., Lin, J., Sjoblom, T., Leary, R. J., et al. (2007). The genomic landscapes of human breast and colorectal cancers. *Science, 318*(5853), 1108–1113.
38. Hulbert, E. M., Smink, L. J., Adlem, E. C., Allen, J. E., Burdick, D. B., Burren, O. S., et al. (2007). T1DBase: Integration and presentation of complex data for type 1 diabetes research. *Nucleic Acids Research, 35*, D742–746.
39. Forbes, S. A., Bhamra, G., Bamford, S., Dawson, E., Kok, C., Clements, J., Menzies, A., Teague, J. W., Futreal, P. A., & Stratton, M. R. (2008). The Catalogue of Somatic Mutations in Cancer (COSMIC). *Current Protococls in Human Genetics*, Chapter 10:Unit 10.11.
40. Church, D. M., Lappalainen, I., Sneddon, T. P., Hinton, J., Maguire, M., Lopez, J., et al. (2010). Public data archives for genomic structural variation. *Nature Genetics, 42*(10), 813–814.
41. Yip, Y. L., Scheib, H., Diemand, A. V., Gattiker, A., Famiglietti, L. M., Gasteiger, E., et al. (2004). The Swiss-Prot variant page and the ModSNP database: A resource for sequence and structure information on human protein variants. *Human Mutation, 23*(5), 464–470.
42. Mottaz, A., David, F. P., Veuthey, A. L., & Yip, Y. L. (2010). Easy retrieval of single amino-acid polymorphisms and phenotype information using SwissVar. *Bioinformatics, 26*(6), 851–852.
43. Cargill, M., Altshuler, D., Ireland, J., Sklar, P., Ardlie, K., Patil, N., et al. (1999). Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nature Genetics, 22*(3), 231–238.
44. Sunyaev, S., Hanke, J., Aydin, A., Wirkner, U., Zastrow, I., Reich, J., et al. (1999). Prediction of nonsynonymous single nucleotide polymorphisms in human disease-associated genes. *Journal of Molecular Medicine, 77*(11), 754–760.
45. Botstein, D., White, R. L., Skolnick, M., & Davis, R. W. (1980). Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *American Journal of Human Genetics, 32*(3), 314–331.
46. Solomon, E., & Bodmer, W. F. (1979). Evolution of sickle variant gene. *Lancet, 1*(8122), 923.
47. Kan, Y. W., & Dozy, A. M. (1978). Polymorphism of DNA sequence adjacent to human beta-globin structural gene: Relationship to sickle mutation. *Proceedings of the National Academy of Sciences of the United States of America, 75*(11), 5631–5635.
48. Feder, J. N., Gnirke, A., Thomas, W., Tsuchihashi, Z., Ruddy, D. A., Basava, A., et al. (1996). A novel MHC class I-like gene is mutated in patients with hereditary haemochromatosis. *Nature Genetics, 13*(4), 399–408.
49. Enattah, N. S., Sahi, T., Savilahti, E., Terwilliger, J. D., Peltonen, L., & Jarvela, I. (2002). Identification of a variant associated with adult-type hypolactasia. *Nature Genetics, 30*(2), 233–237.
50. Kruglyak, L. (2008). The road to genome-wide association studies. *Nat Rev Genet, 9*, 314–318.
51. Sunyaev, S., Ramensky, V., & Bork, P. (2000). Towards a structural basis of human non-synonymous single nucleotide polymorphisms. *Trends in Genetics, 16*(5), 198–200.
52. 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. *Journal of Molecular Biology, 315*(4), 771–786.
53. Ng, P. C., & Henikoff, S. (2001). Predicting deleterious amino acid substitutions. *Genome Research, 11*(5), 863–874.
54. Steward, R. E., MacArthur, M. W., Laskowski, R. A., & Thornton, J. M. (2003). Molecular basis of inherited diseases: A structural perspective. *Trends in Genetics, 19*(9), 505–513.
55. Worth CL, Burke DF, Blundell TL (2007) Estimating the effects of single nucleotide polymorphisms on protein structure: How good are we at identifying likely disease associated mutations? *Proceedings of Molecular Interactions—Bringing Chemistry to Life*, pp. 11–26.
56. Gong, S., Worth, C. L., Bickerton, G. R., Lee, S., Tanramluk, D., & Blundell, T. L. (2009). Structural and functional restraints in the evolution of protein families and superfamilies. *Biochemical Society Transactions, 37*(Pt 4), 727–733.
57. Kimura, M. (1983). *The neutral theory of evolution*. Cambridge: Cambridge University Press.
58. Ohta, T. (1973). Slightly deleterious mutant substitutions in evolution. *Nature, 246*(5428), 96–98.
59. Worth, C. L., Gong, S., & Blundell, T. L. (2009). Structural and functional constraints in the evolution of protein families. *Nature Reviews. Molecular Cell Biology, 10*(10), 709–720.
60. Gusella, J. F., Wexler, N. S., Conneally, P. M., Naylor, S. L., Anderson, M. A., Tanzi, R. E., et al. (1983). A polymorphic DNA marker genetically linked to Huntington's disease. *Nature, 306*(5940), 234–238.
61. Kerem, B., Rommens, J. M., Buchanan, J. A., Markiewicz, D., Cox, T. K., Chakravarti, A., et al. (1989). Identification of the cystic fibrosis gene: Genetic analysis. *Science, 245*(4922), 1073–1080.
62. Riordan, J. R., Rommens, J. M., Kerem, B., Alon, N., Rozmahel, R., Grzelczak, Z., et al. (1989). Identification of the cystic fibrosis gene: Cloning and characterization of complementary DNA. *Science, 245*(4922), 1066–1073.
63. Frazer, K. A., Murray, S. S., Schork, N. J., & Topol, E. J. (2009). Human genetic variation and its contribution to complex traits. *Nature Reviews. Genetics, 10*(4), 241–251.
64. Durbin, R. M., Abecasis, G. R., Altshuler, D. L., Auton, A., Brooks, L. D., Durbin, R. M., et al. (2010). A map of human genome variation from population-scale sequencing. *Nature, 467*(7319), 1061–1073.
65. Matthews, B. W. (1993). Structural and genetic analysis of protein stability. *Annual Review of Biochemistry, 62*, 139–160.
66. Pakula, A. A., & Sauer, R. T. (1989). Genetic analysis of protein stability and function. *Annual Review of Genetics, 23*, 289–310.
67. Ruotolo, B. T., Benesch, J. L., Sandercock, A. M., Hyung, S. J., & Robinson, C. V. (2008). Ion mobility–mass spectrometry analysis of large protein complexes. *Nature Protocols, 3*(7), 1139–1152.
68. McLaughlin, S. H., & Jackson, S. E. (2002). Folding and stability of the ligand-binding domain of the glucocorticoid receptor. *Protein Science, 11*(8), 1926–1936.
69. Perrett, S., Freeman, S. J., Butler, P. J., & Fersht, A. R. (1999). Equilibrium folding properties of the yeast prion protein determinant Ure2. *Journal of Molecular Biology, 290*(1), 331–345.
70. Jackson, S. E., el Masry, N., & Fersht, A. R. (1993). Structure of the hydrophobic core in the transition state for folding of chymotrypsin inhibitor 2: A critical test of the protein engineering method of analysis. *Biochemistry, 32*(42), 11270–11278.
71. Main, E. R., Fulton, K. F., & Jackson, S. E. (1998). Context-dependent nature of destabilizing mutations on the stability of FKBP12. *Biochemistry, 37*(17), 6145–6153.
72. Wray, J. W., Baase, W. A., Lindstrom, J. D., Weaver, L. H., Poteete, A. R., & Matthews, B. W. (1999). Structural analysis of a non-contiguous second-site revertant in T4 lysozyme shows that increasing the rigidity of a protein can enhance its stability. *Journal of Molecular Biology, 292*(5), 1111–1120.
73. Itzhaki, L. S., Evans, P. A., Dobson, C. M., & Radford, S. E. (1994). Tertiary interactions in the folding pathway of hen lysozyme: Kinetic studies using fluorescent probes. *Biochemistry, 33*(17), 5212–5220.

74. Mallam, A. L., & Jackson, S. E. (2007). A comparison of the folding of two knotted proteins: YbeA and YibK. *Journal of Molecular Biology, 366*(2), 650–665.
75. Clarke, J., Hounslow, A. M., & Fersht, A. R. (1995). Disulfide mutants of barnase. II: Changes in structure and local stability identified by hydrogen exchange. *Journal of Molecular Biology, 253*(3), 505–513.
76. Clifford, S. C., Cockman, M. E., Smallwood, A. C., Mole, D. R., Woodward, E. R., Maxwell, P. H., et al. (2001). Contrasting effects on HIF-1alpha regulation by disease-causing pVHL mutations correlate with patterns of tumourigenesis in von Hippel–Lindau disease. *Human Molecular Genetics, 10*(10), 1029–1038.
77. Tanoue, T., Adachi, M., Moriguchi, T., & Nishida, E. (2000). A conserved docking motif in MAP kinases common to substrates, activators and regulators. *Nature Cell Biology, 2*(2), 110–116.
78. Takayama, N., Kizaki, M., Hida, T., Kinjo, K., & Ikeda, Y. (2001). Novel mutation in the PML/RARalpha chimeric gene exhibits dramatically decreased ligand-binding activity and confers acquired resistance to retinoic acid in acute promyelocytic leukemia. *Experimental Hematology, 29*(7), 864–872.
79. Jackson, S. E., & Fersht, A. R. (1994). Contribution of residues in the reactive site loop of chymotrypsin inhibitor two to protein stability and activity. *Biochemistry, 33*(46), 13880–13887.
80. Poliakov, E., Gentleman, S., Cunningham, F. X., Jr., Miller-Ihli, N. J., & Redmond, T. M. (2005). Key role of conserved histidines in recombinant mouse beta-carotene 15,15′-monooxygenase-1 activity. *The Journal of Biological Chemistry, 280*(32), 29217–29223.
81. Alber, T., Sun, D. P., Nye, J. A., Muchmore, D. C., & Matthews, B. W. (1987). Temperature-sensitive mutations of bacteriophage T4 lysozyme occur at sites with low mobility and low solvent accessibility in the folded protein. *Biochemistry, 26*(13), 3754–3758.
82. Clarke, J., Henrick, K., & Fersht, A. R. (1995). Disulfide mutants of barnase. I: Changes in stability and structure assessed by biophysical methods and X-ray crystallography. *Journal of Molecular Biology, 253*(3), 493–504.
83. Matthews, B. W., Nicholson, H., & Becktel, W. J. (1987). Enhanced protein thermostability from site-directed mutations that decrease the entropy of unfolding. *Proceedings of the National Academy of Sciences of the United States of America, 84*(19), 6663–6667.
84. Pace, C. N., Horn, G., Hebert, E. J., Bechert, J., Shaw, K., Urbanikova, L., et al. (2001). Tyrosine hydrogen bonds make a large contribution to protein stability. *Journal of Molecular Biology, 312*(2), 393–404.
85. Stollar, E. J., Mayor, U., Lovell, S. C., Federici, L., Freund, S. M., Fersht, A. R., et al. (2003). Crystal structures of engrailed homeodomain mutants: Implications for stability and dynamics. *The Journal of Biological Chemistry, 278*(44), 43699–43708.
86. Ekblad, C. M., Wilkinson, H. R., Schymkowitz, J. W., Rousseau, F., Freund, S. M., & Itzhaki, L. S. (2002). Characterisation of the BRCT domains of the breast cancer susceptibility gene product BRCA1. *Journal of Molecular Biology, 320*(3), 431–442.
87. Tang, K. S., Guralnick, B. J., Wang, W. K., Fersht, A. R., & Itzhaki, L. S. (1999). Stability and folding of the tumour suppressor protein p16. *Journal of Molecular Biology, 285*(4), 1869–1886.
88. Bullock, A. N., Henckel, J., DeDecker, B. S., Johnson, C. M., Nikolova, P. V., Proctor, M. R., et al. (1997). Thermodynamic stability of wild-type and mutant p53 core domain. *Proceedings of the National Academy of Sciences of the United States of America, 94*(26), 14338–14342.
89. Friedler, A., Veprintsev, D. B., Hansson, L. O., & Fersht, A. R. (2003). Kinetic instability of p53 core domain mutants: Implications for rescue by small molecules. *The Journal of Biological Chemistry, 278*(26), 24108–24112.
90. Nikolova, P. V., Henckel, J., Lane, D. P., & Fersht, A. R. (1998). Semirational design of active tumor suppressor p53 DNA binding domain with enhanced stability. *Proceedings of the National Academy of Sciences of the United States of America, 95*(25), 14675–14680.
91. Joerger, A. C., Allen, M. D., & Fersht, A. R. (2004). Crystal structure of a superstable mutant of human p53 core domain. Insights into the mechanism of rescuing oncogenic mutations. *The Journal of Biological Chemistry, 279*(2), 1291–1296.
92. Joerger, A. C., Ang, H. C., Veprintsev, D. B., Blair, C. M., & Fersht, A. R. (2005). Structures of p53 cancer mutants and mechanism of rescue by second-site suppressor mutations. *The Journal of Biological Chemistry, 280*(16), 16030–16037.
93. Ang, H. C., Joerger, A. C., Mayer, S., & Fersht, A. R. (2006). Effects of common cancer mutations on stability and DNA binding of full-length p53 compared with isolated core domains. *The Journal of Biological Chemistry, 281*(31), 21934–21941.
94. Cheon, D. J., & Orsulic, S. (2011). Mouse models of cancer. *Annu Rev Pathol, 6*, 95–119.
95. Jucker, M. (2010). The benefits and limitations of animal models for translational research in neurodegenerative diseases. *Natural Medicines, 16*(11), 1210–1214.
96. Scheikl, T., Pignolet, B., Mars, L. T., & Liblau, R. S. (2010). Transgenic mouse models of multiple sclerosis. *Cellular and Molecular Life Sciences, 67*(23), 4011–4034.
97. Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., et al. (2001). The sequence of the human genome. *Science, 291*(5507), 1304–1351.
98. Frazer, K. A., Ballinger, D. G., Cox, D. R., Hinds, D. A., Stuve, L. L., Gibbs, R. A., et al. (2007). A second generation human haplotype map of over 3.1 million SNPs. *Nature, 449*(7164), 851–861.
99. Lee, D., Redfern, O., & Orengo, C. (2007). Predicting protein function from sequence and structure. *Nature Reviews. Molecular Cell Biology, 8*(12), 995–1005.
100. Mooney, S. (2005). Bioinformatics approaches and resources for single nucleotide polymorphism functional analysis. *Briefings in Bioinformatics, 6*(1), 44–56.
101. Ng, P. C., & Henikoff, S. (2006). Predicting the effects of amino acid substitutions on protein function. *Annual Review of Genomics and Human Genetics, 7*, 61–80.
102. Topham, C. M., Srinivasan, N., & Blundell, T. L. (1997). Prediction of the stability of protein mutants based on structural environment-dependent amino acid substitution and propensity tables. *Protein Engineering, 10*(1), 7–21.
103. Guerois, R., Nielsen, J. E., & Serrano, L. (2002). Predicting changes in the stability of proteins and protein complexes: A study of more than 1000 mutations. *Journal of Molecular Biology, 320*(2), 369–387.
104. Capriotti, E., Fariselli, P., & Casadio, R. (2004). A neural-network-based method for predicting protein stability changes upon single point mutations. *Bioinformatics, 20*(1), i63–i68.
105. Capriotti, E., Fariselli, P., Calabrese, R., & Casadio, R. (2005). Predicting protein stability changes from sequences using support vector machines. *Bioinformatics, 21*(2), 54–58.
106. Capriotti, E., Fariselli, P., & Casadio, R. (2005). I-Mutant2.0: Predicting stability changes upon mutation from the protein sequence or structure. *Nucleic Acids Research, 33*, W306–310.
107. Cheng, J., Randall, A., & Baldi, P. (2006). Prediction of protein stability changes for single-site mutations using support vector machines. *Proteins, 62*(4), 1125–1132.
108. Parthiban, V., Gromiha, M. M., & Schomburg, D. (2006). CUPSAT: Prediction of protein stability upon point mutations. *Nucleic Acids Research, 34*, W239–242.
109. Yin, S., Ding, F., & Dokholyan, N. V. (2007). Modeling backbone flexibility improves protein stability estimation. *Structure, 15*(12), 1567–1576.

110. Fernandez-Escamilla, A. M., Rousseau, F., Schymkowitz, J., & Serrano, L. (2004). Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins. *Nature Biotechnology, 22*(10), 1302–1306.

111. Conchillo-Sole, O., de Groot, N. S., Aviles, F. X., 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.

112. Trovato, A., Seno, F., & Tosatto, S. C. (2007). The PASTA server for protein aggregation prediction. *Protein Engineering, Design & Selection, 20*(10), 521–523.

113. Morgan, D. H., Kristensen, D. M., Mittelman, D., & Lichtarge, O. (2006). ET viewer: An application for predicting and visualizing functional sites in protein structures. *Bioinformatics, 22*(16), 2049–2050.

114. Joachimiak, M. P., & Cohen, F. E. (2002). JEvTrace: Refinement and variations of the evolutionary trace in JAVA. *Genome Biology, 3*(12), RESEARCH0077.

115. La, D., & Livesay, D. R. (2005). MINER: Software for phylogenetic motif identification. *Nucleic Acids Research, 33*, W267–270.

116. Chelliah, V., Blundell, T., & Mizuguchi, K. (2005). Functional restraints on the patterns of amino acid substitutions: Application to sequence–structure homology recognition. *Proteins, 61*(4), 722–731.

117. Porter, C. T., Bartlett, G. J., & Thornton, J. M. (2004). The Catalytic Site Atlas: A resource of catalytic sites and residues identified in enzymes using structural data. *Nucleic Acids Research, 32*, D129–133.

118. Ivanisenko, V. A., Pintus, S. S., Grigorovich, D. A., & Kolchanov, N. A. (2004). PDBSiteScan: A program for searching for active, binding and posttranslational modification sites in the 3D structures of proteins. *Nucleic Acids Research, 32*, W549–554.

119. Golovin, A., Dimitropoulos, D., Oldfield, T., Rachedi, A., & Henrick, K. (2005). MSDsite: A database search and retrieval system for the analysis and viewing of bound ligands and active sites. *Proteins, 58*(1), 190–199.

120. Rohl, C. A., Strauss, C. E., Misura, K. M., & Baker, D. (2004). Protein structure prediction using Rosetta. *Methods Enzymol, 383*, 66–93.

121. Ng, P. C., & Henikoff, S. (2002). Accounting for human polymorphisms predicted to affect protein function. *Genome Research, 12*(3), 436–446.

122. Sunyaev, S., Ramensky, V., Koch, I., Lathe, W., 3rd, Kondrashov, A. S., & Bork, P. (2001). Prediction of deleterious human alleles. *Human Molecular Genetics, 10*(6), 591–597.

123. Bao, L., Zhou, M., & Cui, Y. (2005). nsSNPAnalyzer: Identifying disease-associated nonsynonymous single nucleotide polymorphisms. *Nucleic Acids Research, 33*, W480–482.

124. Bromberg, Y., & Rost, B. (2007). SNAP: Predict effect of non-synonymous polymorphisms on function. *Nucleic Acids Research, 35*(11), 3823–3835.

125. 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*(22), 2729–2734.

126. Blundell, T. L., Cooper, J., Donnelly, D., Driessen, H., Edwards, Y., Eisenmenger, F., et al. (1991). Patterns of sequence variation in families of homologous proteins. In H. Jornvall, J. O. Hoog, & A. M. Gustavsson (Eds.), *Methods in proteins sequence analysis* (pp. 373–385). Basel: Birkhauser Verlag AG.

127. Overington, J., Johnson, M. S., Sali, A., & Blundell, T. L. (1990). Tertiary structural constraints on protein evolutionary diversity: Templates, key residues and structure prediction. *Proc Biol Sci, 241*(1301), 132–145.

128. Ferguson, B. J., Alexander, C., Rossi, S. W., Liiv, I., Rebane, A., Worth, C. L., et al. (2008). AIRE’s CARD revealed, a new structure for central tolerance provokes transcriptional plasticity. *The Journal of Biological Chemistry, 283*(3), 1723–1731.

129. Velho, S., Oliveira, C., Paredes, J., Sousa, S., Leite, M., Matos, P., et al. (2010). Mixed lineage kinase three gene mutations in mismatch repair deficient gastrointestinal tumours. *Human Molecular Genetics, 19*(4), 697–706.

130. Nagpal, K., Plantinga, T. S., Wong, J., Monks, B. G., Gay, N. J., Netea, M. G., et al. (2009). A TIR domain variant of MyD88 adapter-like (Mal)/TIRAP results in loss of MyD88 binding and reduced TLR2/TLR4 signaling. *The Journal of Biological Chemistry, 284*(38), 25742–25748.

131. Rowling, P. J., Cook, R., & Itzhaki, L. S. (2010). Toward classification of BRCA1 missense variants using a biophysical approach. *The Journal of Biological Chemistry, 285*(26), 20080–20087.

132. Chelliah, V., Chen, L., Blundell, T. L., & Lovell, S. C. (2004). Distinguishing structural and functional restraints in evolution in order to identify interaction sites. *Journal of Molecular Biology, 342*(5), 1487–1504.

133. Lee, S., & Blundell, T. L. (2009). BIPA: A database for protein–nucleic acid interaction in 3D structures. *Bioinformatics, 25*(12), 1559–1560.

134. Schreyer, A., & Blundell, T. L. (2009). A protein–ligand interaction database for drug discovery. *Chemical Biology & Drug Design, 73*, 157–167.

135. Forman, J. R., Worth, C. L., Bickerton, G. R., Eisen, T. G., & Blundell, T. L. (2009). Structural bioinformatics mutation analysis reveals genotype–phenotype correlations in von Hippel–Lindau disease and suggests molecular mechanisms of tumorigenesis. *Proteins, 77*(1), 84–96.

136. Cangul, H., Morgan, N. V., Forman, J. R., Saglam, H., Aycan, Z., Yakut, T., et al. (2010). Novel TSHR mutations in consanguineous families with congenital nongoitrous hypothyroidism. *Clin Endocrinol (Oxf), 73*(5), 671–677.

137. Ricketts, C. J., Forman, J. R., Rattenberry, E., Bradshaw, N., Lalloo, F., Izatt, L., et al. (2010). Tumor risks and genotype–phenotype–proteotype analysis in 358 patients with germline mutations in SDHB and SDHD. *Human Mutation, 31*(1), 41–51.

138. Cheng, T. M., Lu, Y. E., Vendruscolo, M., Lio, P., & Blundell, T. L. (2008). Prediction by graph theoretic measures of structural effects in proteins arising from non-synonymous single nucleotide polymorphisms. *PLoS Computational Biology, 4*(7), e1000135.

139. Thomas, P. D., Campbell, M. J., Kejariwal, A., Mi, H., Karlak, B., Daverman, R., et al. (2003). PANTHER: A library of protein families and subfamilies indexed by function. *Genome Research, 13*(9), 2129–2141.

140. Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., et al. (2010). A method and server for predicting damaging missense mutations. *Nat Methods, 7*(4), 248–249.

141. Bickerton, G. R. (2009). *Molecular characterization and evolutionary plasticity of protein–protein interfaces*. Cambridge: Emmanuel College, University of Cambridge.

142. Lee, S., Brown, A., Pitt, W. R., Perez Higueruelo, A., Gong, S., Bickerton, G. R., et al. (2009). Structural interactomics: Informatics approaches to aid the interpretation of genetic variation and the development of novel therapeutics. *Molecular Biosystems, 5*, 1456–1472.

143. Mizuguchi, K., Deane, C. M., Blundell, T. L., Johnson, M. S., & Overington, J. P. (1998). JOY: Protein sequence–structure representation and analysis. *Bioinformatics, 14*(7), 617–623.

144. Jmol: An open-source Java viewer for chemical structures in 3D. http://www.jmol.org/.

145. Hubbard, T. J., Aken, B. L., Ayling, S., Ballester, B., Beal, K., Bragin, E., et al. (2009). Ensembl 2009. *Nucleic Acids Research, 37*, D690–697.
146. Yip, Y. L., Famiglietti, M., Gos, A., Duek, P. D., David, F. P., Gateau, A., et al. (2008). Annotating single amino acid polymorphisms in the UniProt/Swiss-Prot knowledgebase. *Human Mutation, 29*(3), 361–366.
147. Gong, S., & Blundell, T. L. (2008). Discarding functional residues from the substitution table improves predictions of active sites within three-dimensional structures. *PLoS Computational Biology, 4*(10), e1000179.
148. Stein, L. D., Mungall, C., Shu, S., Caudy, M., Mangone, M., Day, A., et al. (2002). The generic genome browser: A building block for a model organism system database. *Genome Research, 12*(10), 1599–1610.
149. Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., et al. (2002). The human genome browser at UCSC. *Genome Research, 12*(6), 996–1006.
150. Harris, T. W., Antoshechkin, I., Bieri, T., Blasiar, D., Chan, J., Chen, W. J., et al. (2010). WormBase: A comprehensive resource for nematode research. *Nucleic Acids Research, 38*, D463–467.
151. Prlic, A., Down, T. A., Kulesha, E., Finn, R. D., Kahari, A., & Hubbard, T. J. (2007). Integrating sequence and structural biology with DAS. *BMC Bioinformatics, 8*, 333.
152. Sanger, F., & Tuppy, H. (1951). The amino-acid sequence in the phenylalanyl chain of insulin. 2. The investigation of peptides from enzymic hydrolysates. *The Biochemical Journal, 49*(4), 481–490.
153. Sanger, F., & Tuppy, H. (1951). The amino-acid sequence in the phenylalanyl chain of insulin I. The identification of lower peptides from partial hydrolysates. *The Biochemical Journal, 49*(4), 463–481.
154. Consortium TU. (2007). The Universal Protein Resource (UniProt). *Nucleic Acids Research, 35*, D193–197.
155. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., et al. (2000). The Protein Data Bank. *Nucleic Acids Research, 28*(1), 235–242.
156. Laskowski, R. A., & Thornton, J. M. (2008). Understanding the molecular machinery of genetics through 3D structures. *Nature Reviews. Genetics, 9*(2), 141–151.
157. Sali, A., & Blundell, T. L. (1990). Definition of general topological equivalence in protein structures. A procedure involving comparison of properties and relationships through simulated annealing and dynamic programming. *Journal of Molecular Biology, 212*(2), 403–428.
158. Sali, A., Overington, J. P., Johnson, M. S., & Blundell, T. L. (1990). From comparisons of protein sequences and structures to protein modelling and design. *Trends in Biochemical Sciences, 15*(6), 235–240.
159. Moult, J. (2005). A decade of CASP: Progress, bottlenecks and prognosis in protein structure prediction. *Current Opinion in Structural Biology, 15*(3), 285–289.
160. Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology, 215*(3), 403–410.
161. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., et al. (1997). Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Research, 25*(17), 3389–3402.
162. Finn, R. D., Tate, J., Mistry, J., Coggill, P. C., Sammut, S. J., Hotz, H. R., et al. (2008). The Pfam protein families database. *Nucleic Acids Research, 36*, D281–288.
163. Rost, B. (1995). TOPITS: Threading one-dimensional predictions into three-dimensional structures. *Proc Int Conf Intell Syst Mol Biol, 3*, 314–321.
164. Shi, J., Blundell, T. L., & Mizuguchi, K. (2001). FUGUE: Sequence–structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. *Journal of Molecular Biology, 310*(1), 243–257.
165. Furnham, N., de Bakker, P. I., Gore, S., Burke, D. F., & Blundell, T. L. (2008). Comparative modelling by restraint-based conformational sampling. *BMC Structural Biology, 8*(1), 7.
166. Gore, S. P., Karmali, A. M., & Blundell, T. L. (2007). Rappertk: A versatile engine for discrete restraint-based conformational sampling of macromolecules. *BMC Structural Biology, 7*, 13.
167. Sali, A., & Blundell, T. L. (1993). Comparative protein modelling by satisfaction of spatial restraints. *Journal of Molecular Biology, 234*(3), 779–815.
168. Bates, P. A., Kelley, L. A., MacCallum, R. M., & Sternberg, M. J. (2001). Enhancement of protein modeling by human intervention in applying the automatic programs 3D-JIGSAW and 3D-PSSM. *Proteins Suppl, 5*, 39–46.
169. Montalvao, R. W., Smith, R. E., Lovell, S. C., & Blundell, T. L. (2005). CHORAL: A differential geometry approach to the prediction of the cores of protein structures. *Bioinformatics, 21*(19), 3719–3725.
170. Peitsch, M. C., Wilkins, M. R., Tonella, L., Sanchez, J. C., Appel, R. D., & Hochstrasser, D. F. (1997). Large-scale protein modelling and integration with the SWISS-PROT and SWISS-2DPAGE databases: The example of *Escherichia coli*. *Electrophoresis, 18*(3–4), 498–501.
171. Sutcliffe, M. J., Hayes, F. R., & Blundell, T. L. (1987). Knowledge based modelling of homologous proteins, part II: Rules for the conformations of substituted sidechains. *Protein Engineering, 1*(5), 385–392.
172. Lovell, S. C., Davis, I. W., Arendall, W. B., 3rd, de Bakker, P. I., Word, J. M., Prisant, M. G., et al. (2003). Structure validation by Calpha geometry: Phi, psi and Cbeta deviation. *Proteins, 50*(3), 437–450.
173. Sippl, M. J. (1993). Recognition of errors in three-dimensional structures of proteins. *Proteins, 17*(4), 355–362.
174. Bradley, P., Misura, K. M., & Baker, D. (2005). Toward high-resolution de novo structure prediction for small proteins. *Science, 309*(5742), 1868–1871.
175. Alimonti, A., Carracedo, A., Clohessy, J. G., Trotman, L. C., Nardella, C., Egia, A., et al. (2010). Subtle variations in Pten dose determine cancer susceptibility. *Nature Genetics, 42*(5), 454–458.
176. Hindorff, L. A., Sethupathy, P., Junkins, H. A., Ramos, E. M., Mehta, J. P., Collins, F. S., et al. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences of the United States of America, 106*(23), 9362–9367.
177. Chamary, J. V., Parmley, J. L., & Hurst, L. D. (2006). Hearing silence: Non-neutral evolution at synonymous sites in mammals. *Nature Rev Genet, 7*, 98–108.
178. Clark, T. G., Andrew, T., Cooper, G. M., Margulies, E. H., Mullikin, J. C., & Balding, D. J. (2007). Functional constraint and small insertions and deletions in the ENCODE regions of the human genome. *Genome Biology, 8*(9), R180.
179. Mills, R. E., Luttig, C. T., Larkins, C. E., Beauchamp, A., Tsui, C., Pittard, W. S., et al. (2006). An initial map of insertion and deletion (INDEL) variation in the human genome. *Genome Research, 16*(9), 1182–1190.
180. Redon, R., Ishikawa, S., Fitch, K. R., Feuk, L., Perry, G. H., Andrews, T. D., et al. (2006). Global variation in copy number in the human genome. *Nature, 444*(7118), 444–454.
181. Sebat, J., Lakshmi, B., Troge, J., Alexander, J., Young, J., Lundin, P., et al. (2004). Large-scale copy number polymorphism in the human genome. *Science, 305*(5683), 525–528.
182. Sudmant, P. H., Kitzman, J. O., Antonacci, F., Alkan, C., Malig, M., Tsalenko, A., et al. (2010). Diversity of human copy number variation and multicopy genes. *Science, 330*(6004), 641–646.

183. Gemayel, R., Vinces, M. D., Legendre, M., & Verstrepen, K. J. (2010). Variable tandem repeats accelerate evolution of coding and regulatory sequences. *Annual Review of Genetics, 44*, 445–477.

184. McCarroll, S. A. (2010). Copy number variation and human genome maps. *Nature Genetics, 42*(5), 365–366.

185. Mullaney, J. M., Mills, R. E., Pittard, W. S., & Devine, S. E. (2010). Small insertions and deletions (INDELs) in human genomes. *Human Molecular Genetics, 19*(2), R131–136.

186. Soskine, M., & Tawfik, D. S. (2010). Mutational effects and the evolution of new protein functions. *Nature Reviews. Genetics, 11* (8), 572–582.

187. Stankiewicz, P., & Lupski, J. R. (2010). Structural variation in the human genome and its role in disease. *Annual Review of Medicine, 61*, 437–455.

188. Wain, L. V., Armour, J. A., & Tobin, M. D. (2009). Genomic copy number variation, human health, and disease. *Lancet, 374* (9686), 340–350.

189. Goldberg, A. L. (2003). Protein degradation and protection against misfolded or damaged proteins. *Nature, 426*(6968), 895–899.

190. Ferrer-Costa, C., Orozco, M., & de la Cruz, X. (2007). Characterization of compensated mutations in terms of structural and physico-chemical properties. *Journal of Molecular Biology, 365*(1), 249–256.

191. Marguerat, S., Wilhelm, B. T., & Bahler, J. (2008). Next-generation sequencing: Applications beyond genomes. *Biochemical Society Transactions, 36*(Pt 5), 1091–1096.

192. Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: A revolutionary tool for transcriptomics. *Nature Reviews. Genetics, 10*(1), 57–63.

193. Brookes, A. J., Lehvaslaiho, H., Siegfried, M., Boehm, J. G., Yuan, Y. P., Sarkar, C. M., et al. (2000). HGBASE: A database of SNPs and other variations in and around human genes. *Nucleic Acids Research, 28*(1), 356–360.

194. Fredman, D., Siegfried, M., Yuan, Y. P., Bork, P., Lehvaslaiho, H., & Brookes, A. J. (2002). HGVbase: A human sequence variation database emphasizing data quality and a broad spectrum of data sources. *Nucleic Acids Research, 30*(1), 387–391.

195. Gromiha, M. M., An, J., Kono, H., Oobatake, M., Uedaira, H., & Sarai, A. (1999). ProTherm: Thermodynamic Database for Proteins and Mutants. *Nucleic Acids Research, 27*(1), 286–288.

196. Thorn, K. S., & Bogan, A. A. (2001). ASEdb: A database of alanine mutations and their effects on the free energy of binding in protein interactions. *Bioinformatics, 17*(3), 284–285.

197. Martin, A. C., Facchiano, A. M., Cuff, A. L., Hernandez-Boussard, T., Olivier, M., Hainaut, P., et al. (2002). Integrating mutation data and structural analysis of the TP53 tumor-suppressor protein. *Human Mutation, 19*(2), 149–164.

198. Kwok, C. J., Martin, A. C., Au, S. W., & Lam, V. M. (2002). G6PDdb, an integrated database of glucose-6-phosphate dehydrogenase (G6PD) mutations. *Human Mutation, 19*(3), 217–224.

199. Mooney, S. D., & Altman, R. B. (2003). MutDB: Annotating human variation with functionally relevant data. *Bioinformatics, 19*(14), 1858–1860.

200. Riva, A., & Kohane, I. S. (2002). SNPper: Retrieval and analysis of human SNPs. *Bioinformatics, 18*(12), 1681–1685.

201. Bamford, S., Dawson, E., Forbes, S., Clements, J., Pettett, R., Dogan, A., et al. (2004). The COSMIC (Catalogue of Somatic Mutations in Cancer) database and website. *British Journal of Cancer, 91*(2), 355–358.

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

203. Karchin, R., Diekhans, M., Kelly, L., Thomas, D. J., Pieper, U., Eswar, N., et al. (2005). LS-SNP: Large-scale annotation of coding non-synonymous SNPs based on multiple information sources. *Bioinformatics, 21*(12), 2814–2820.

204. Hurst, J. M., McMillan, L. E., Porter, C. T., Allen, J., Fakorede, A., & Martin, A. C. (2009). The SAAPdb web resource: A large-scale structural analysis of mutant proteins. *Human Mutation, 30* (4), 616–624.

205. Reumers, J., Maurer-Stroh, S., Schymkowitz, J., & Rousseau, F. (2006). SNPeffect v2.0: A new step in investigating the molecular phenotypic effects of human non-synonymous SNPs. *Bioinformatics, 22*(17), 2183–2185.

206. Reumers, J., Schymkowitz, J., Ferkinghoff-Borg, J., Stricher, F., Serrano, L., & Rousseau, F. (2005). SNPeffect: A database mapping molecular phenotypic effects of human non-synonymous coding SNPs. *Nucleic Acids Research, 33*, D527–532.

207. Han, A., Kang, H. J., Cho, Y., Lee, S., Kim, Y. J., & Gong, S. (2006). SNP@Domain: A web resource of single nucleotide polymorphisms (SNPs) within protein domain structures and sequences. *Nucleic Acids Research, 34*, W642–644.

208. Jegga, A. G., Gowrisankar, S., Chen, J., & Aronow, B. J. (2007). PolyDoms: A whole genome database for the identification of non-synonymous coding SNPs with the potential to impact disease. *Nucleic Acids Research, 35*, D700–706.

209. Peterson, T. A., Adadey, A., Santana-Cruz, I., Sun, Y., Winder, A., & Kann, M. G. (2010). DMDM: Domain mapping of disease mutations. *Bioinformatics, 26*(19), 2458–2459.

210. Craddock, N., Hurles, M. E., Cardin, N., Pearson, R. D., Plagnol, V., Robson, S., et al. (2010). Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. *Nature, 464*(7289), 713–720.

211. Topham, C. M., McLeod, A., Eisenmenger, F., Overington, J. P., Johnson, M. S., & Blundell, T. L. (1993). Fragment ranking in modelling of protein structure. Conformationally constrained environmental amino acid substitution tables. *Journal of Molecular Biology, 229*(1), 194–220.

212. Dehouck, Y., Grosfils, A., Folch, B., Gilis, D., Bogaerts, P., & Rooman, M. (2009). Fast and accurate predictions of protein stability changes upon mutations using statistical potentials and neural networks: PoPMuSiC-2.0. *Bioinformatics, 25*, 2537–2543.

213. Gilis, D., & Rooman, M. (2000). PoPMuSiC, an algorithm for predicting protein mutant stability changes: Application to prion proteins. *Protein Engineering, 13*(12), 849–856.

214. 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 Science, 11*(11), 2714–2726.

215. Schymkowitz, J., Borg, J., Stricher, F., Nys, R., Rousseau, F., & Serrano, L. (2005). The FoldX web server: An online force field. *Nucleic Acids Research, 33*, W382–388.

216. Ramensky, V., Bork, P., & Sunyaev, S. (2002). Human non-synonymous SNPs: Server and survey. *Nucleic Acids Research, 30*(17), 3894–3900.

217. Christen, M., Hunenberger, P. H., Bakowies, D., Baron, R., Burgi, R., Geerke, D. P., et al. (2005). The GROMOS software for biomolecular simulation: GROMOS05. *Journal of Computational Chemistry, 26*(16), 1719–1751.

218. Ferrer-Costa, C., Gelpi, J. L., 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*(14), 3176–3178.

219. Yuan, H. Y., Chiou, J. J., Tseng, W. H., Liu, C. H., Liu, C. K., Lin, Y. J., et al. (2006). FASTSNP: An always up-to-date and extendable service for SNP function analysis and prioritization. *Nucleic Acids Research, 34*, W635–641.

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

221. Yin, S., Ding, F., & Dokholyan, N. V. (2007). Eris: An automated estimator of protein stability. *Nat Methods, 4*(6), 466–467.

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

223. Uzun, A., Leslin, C. M., Abyzov, A., & Ilyin, V. (2007). Structure SNP (StSNP): A web server for mapping and modeling nsSNPs on protein structures with linkage to metabolic pathways. *Nucleic Acids Research, 35*, W384–392.

224. Li, S., Ma, L., Li, H., Vang, S., Hu, Y., Bolund, L., et al. (2007). Snap: An integrated SNP annotation platform. *Nucleic Acids Research, 35*, D707–710.

225. Masso, M., & Vaisman, I. I. (2010). AUTO-MUTE: Web-based tools for predicting stability changes in proteins due to single amino acid replacements. *Protein Engineering, Design & Selection, 23*(8), 683–687.

226. Capriotti, E., Arbiza, L., Casadio, R., Dopazo, J., Dopazo, H., & Marti-Renom, M. A. (2008). Use of estimated evolutionary strength at the codon level improves the prediction of disease-related protein mutations in humans. *Human Mutation, 29*(1), 198–204.

227. Lee, P. H., & Shatkay, H. (2008). F-SNP: Computationally predicted functional SNPs for disease association studies. *Nucleic Acids Research, 36*, D820–824.

228. Brooks, B. R., Brooks, C. L., 3rd, Mackerell, A. D., Jr., Nilsson, L., Petrella, R. J., Roux, B., et al. (2009). CHARMM: The biomolecular simulation program. *Journal of Computational Chemistry, 30*(10), 1545–1614.

229. Kotowski, I. K., Pertsemlidis, A., Luke, A., Cooper, R. S., Vega, G. L., Cohen, J. C., et al. (2006). A spectrum of PCSK9 alleles contributes to plasma levels of low-density lipoprotein cholesterol. *American Journal of Human Genetics, 78*(3), 410–422.

230. Allard, D., Amsellem, S., Abifadel, M., Trillard, M., Devillers, M., Luc, G., et al. (2005). Novel mutations of the PCSK9 gene cause variable phenotype of autosomal dominant hypercholesterolemia. *Human Mutation, 26*(5), 497.

231. Murzin, A. G., Brenner, S. E., Hubbard, T., & Chothia, C. (1995). SCOP: A structural classification of proteins database for the investigation of sequences and structures. *Journal of Molecular Biology, 247*(4), 536–540.

AUTHOR QUERY

AUTHOR PLEASE ANSWER QUERY.

Q1. [Experimental Techniques for Investigating Effect(s) of Mutations on Proteins]"…can also be investigated using fluorescence and far-UV CD measurements, where folding is reversible the renaturation curve will superimpose on the denaturation curve" was changed to "…can also be investigated using fluorescence and far-UV CD measurements where folding is reversible and the renaturation curve will superimpose on the denaturation curve". Please check.