Modelling the effects of single point mutations on the structure and function of proteins

Inauguraldissertation
zur
Erlangung der Würde eines Doktors der Philosophie
vorgelegt der
Philosophisch-Naturwissenschaftlichen Fakultät der
Universität Basel

von

James Battey

aus

Grossbritannien

Basel, 2009

Genehmigt von der Philosophisch-Naturwissenschaftlichen Fakultät auf Antrag von

Prof. Dr. Torsten Schwede

Prof. Dr. Olivier Michielin

Basel, den 09.12.2008

Prof. Dr. Eberhard Parlow

Dekan

Namensnennung-Keine kommerzielle Nutzung-Keine Bearbeitung 2.5 Schweiz

Sie dürfen:

das Werk vervielfältigen, verbreiten und öffentlich zugänglich machen

Zu den folgenden Bedingungen:

Namensnennung. Sie müssen den Namen des Autors/Rechteinhabers in der von ihm festgelegten Weise nennen (wodurch aber nicht der Eindruck entstehen darf, Sie oder die Nutzung des Werkes durch Sie würden entlohnt).

Keine kommerzielle Nutzung. Dieses Werk darf nicht für kommerzielle Zwecke verwendet werden.

Keine Bearbeitung. Dieses Werk darf nicht bearbeitet oder in anderer Weise verändert werden.

- Im Falle einer Verbreitung müssen Sie anderen die Lizenzbedingungen, unter welche dieses Werk fällt, mitteilen. Am Einfachsten ist es, einen Link auf diese Seite einzubinden.

- Jede der vorgenannten Bedingungen kann aufgehoben werden, sofern Sie die Einwilligung des Rechteinhabers dazu erhalten.
- Diese Lizenz lässt die Urheberpersönlichkeitsrechte unberührt.

Die gesetzlichen Schranken des Urheberrechts bleiben hiervon unberührt.

Die Commons Deed ist eine Zusammenfassung des Lizenzvertrags in allgemeinverständlicher Sprache: http://creativecommons.org/licenses/by-nc-nd/2.5/ch/legalcode.de

Haftungsausschluss:
Die Commons Deed ist kein Lizenzvertrag. Sie ist lediglich ein Referenztext, der den zugrundeliegenden Lizenzvertrag übersichtlich und in allgemeinverständlicher Sprache wiedergibt. Die Deed selbst entfaltet keine juristische Wirkung und erscheint im eigentlichen Lizenzvertrag nicht. Creative Commons ist keine Rechtsanwaltsgesellschaft und leistet keine Rechtsberatung. Die Weitergabe und Verlinkung des Commons Deeds führt zu keinem Mandatsverhältnis.

Quelle: http://creativecommons.org/licenses/by-nc-nd/2.5/ch/ Datum: 3.4.2009

Abstract

Insight into the molecular impact of mutations on the structure and function of proteins is of great importance in biology. It helps understand the evolution of proteins, rationalize the molecular causes of disease and, from a practical perspective, aid in planning experiments. In this work, three goals are pursued. Firstly, a method for objectively assessing the effect of mutations on protein structure is formulated. The random noise component in the comparison of two structures is quantified by a log linear regression model incorporating information on experimental quality and intrinsic flexibility, which can account for approximately half of all structural variation between alternative structures. Applying this model to the task of isolating the effects of single point mutations, it is shown that subtle changes in structure, induced by mutations from evolutionarily favourable residues to unfavourable ones, can't be observed without correcting for noise. Secondly, the use of automated prediction tools for generating 3D structures for proteins without experimental structures is assessed. It is found that current state of the art automated modelling methods rival or exceed most expert modelling groups in terms of coverage and accuracy. However, in both cases there is still significant room for improvement until protein structure model reach accuracy comparable to experimental structures for non-trivial target proteins. Computationally cheap methods fare comparatively well and thus represent useful tools for the purpose of providing valuable structural information for systematic analyses, such as the study at hand. Finally, the use of machine learning methods for predicting the impact of mutations on protein function is assessed, using a large set of single amino acid variants in humans. The contribution of structural and evolutionary information to predicting the phenotype of mutations is tested rigorously and it is found that structural information provides information not present in evolutionary data. A generalised classifier using both sequence and structure derived information outperforms other comparable published methods. By validating the classifier on independent datasets we show that it can be used as a general purpose mutation prediction tool, and that our validation methods give reasonable estimates of its performance.

Contents

1. Introduction ... 1

1.1 Proteins and their mutability ... 1

1.1.1 Proteins and their role in organisms ... 1

1.1.2 Effects of mutations on proteins ... 6

1.1.3 Relevance of point mutation effect prediction ... 9

1.2 Modelling the effects of mutations ... 11

1.2.1 Impact on protein structure ... 11

1.2.2 Impact on protein function ... 14

1.3 Problem definition ... 25

Modelling the structural impact of mutations ... 25

Extending structural coverage through comparative modelling ... 25

Approaches to improving phenotype prediction ... 26

2. Results and Discussion ... 29

2.1. Assessing the structural impact of mutations ... 29

Prediction of the random component of variability in protein X-ray structures and its implication for interpreting the effects of single point mutations ... 29

2.2 Assessment of template-based predictions in the CASP7 experiment ... 63

2.3: Predicting the functional impact of mutations ... 101

Predicting the functional impact of single point mutations: a test of accuracy and robustness of descriptors and machine learning methods. ... 101

3. Summary and Outlook ... 135

4. References ... 139

5. Acknowledgements ... 147

6. Curriculum Vitae ... 149

1. Introduction

1.1 Proteins and their mutability

1.1.1 Proteins and their role in organisms

Proteins are important, biologically active molecules in all living organisms. Their amino acid sequence is encoded in the DNA, and the ultimate purpose of proteins is to ensure the survival and replication thereof. Their function and biophysical properties are briefly described in the following section.

Proteins as workhorses of the cell

Proteins are the molecular effectors of the life. They perform a large proportion of the functions necessary to the growth, survival and ultimately replication of cells and organisms. They are involved in almost every aspect of the life of an organism or cell, ranging from the replication of life's template, DNA, over the production of energy and the chemical building blocks of life, to regulating a cell's growth and its function within an organism. Proteins are involved in diverse tasks, which can be of catalytic (tasked with replication, metabolism and cell signalling) or structural nature (cytoskeletal proteins); they are also involved in transport of substances throughout an organism (haemoglobin, LDL/HDL) or across membrane barriers (acquaporin, glucose permease), as well as immunity (immunoglobulins).

The importance of protein structure

The function of a protein is intimately linked to its structural properties. For instance, the residues of active sites in enzymes need to be in a specific three-dimensional arrangement in order to catalyse reactions, as exemplified by the catalytic triad in serine proteases. Binding pockets of active sites need a certain shape in order to guarantee ligand specificity. The shape and surface properties are important for specific binding to other proteins in an organised and highly controlled fashion.

The distinctive three dimensional shape of a protein is determined by its amino acid sequence, meaning that the information required for a protein is encoded in the linear amino acid sequence. Even upon denaturation many globular proteins will refold to their native state once removed from the denaturing conditions. This principle, named Anfinsen's paradigm after its discoverer[1], has pervaded the understanding of protein folding ever since its inception and appears to be applicable for practically all known proteins. Protein folds, i.e. the relative three dimensional arrangement of secondary structure elements, are thought to be robust to sequence change during evolution. Naturally occurring proteins similar in sequence are generally also similar in structure, and only proteins very different in sequence tend to, but need not, be structurally distinct[2].

While these principles hold for most known proteins, there are some exceptions, which merit brief discussion[3]. The recent discovery of two distinct folds for lymphotactin[4], which are assumed at different salt concentrations and temperature conditions, shows that it is possible for one sequence to undergo not just local restructuring, but global restructuring to the extent that the two folds have no hydrogen bonds in common. A further example is the class of the cysteine rich domains of cnidarians nematocyst proteins, which can natively interconvert between folds, each of which can be stabilized by introducing very few mutations[5]. The accepted paradigm, which posits that proteins similar in sequence also have similar folds, appears not to hold for such proteins; as such these proteins are assumed to represent evolutionary bridges between folds which permit the exploration of structure space. More notably, there are proteins which are not intrinsically structured. Proteins which are natively disordered, i.e. they do not assume a clear structural fold, have recently gained much interest[6,7]. This intrinsic disorder appears to be a property of many proteins involved in important cell functions such as signalling, development and cell cycle control[8]. It is however assumed that these proteins undergo structuring upon forming complexes with their binding partners[9].

Determinants of protein folding and stability

Upon transcription of the mRNA to proteins by the ribosome, proteins exist initially as an unstructured linear sequence of amino acids. The process it undergoes to obtain its three dimensional structure is known as protein folding. The number of available conformations that a protein can assume is in principle astronomic. Under a very simplified model, a single amino acid can assume three conformations, which is an underestimation of the true degrees of freedom, a peptide of length N can access 3^N different conformations. For even small proteins, sampling all of these conformations is impossible to achieve in the amount of time a protein has to fold. This apparent contradiction implies that in order to fold in a reasonable time, kinetic pathways or protein folding pathways have to exist, which a protein can follow in order to quickly reach its native state[10]. Much debate exists as to the exact mechanisms of such pathways[11], but it is clear that they must be encoded for in the amino acid sequence of the protein at hand.

The stability of a protein is defined by the ratio of folded of unfolded copies of the protein in a system. As such, a protein's stability determines its effective concentration in the cell, i.e. the concentration of folded and therefore functional molecules. Generally speaking, the native fold of most proteins are only transiently stable and for many known proteins is about -3 to -14 kcal/mol[12], which is on the order of magnitude of the energetic contribution of a single hydrogen bond. This transient stability makes proteins susceptible to unfolding, or denaturation, by heat and chemical effects such as strong changes in pH or high concentrations of denaturants. Under physiological conditions however, the thermodynamic balance drives proteins to their unique folded state. Selection appears to maintain a minimum stability level for protein, so as to allow it to adopt its functional form long enough to perform its function. Insights from protein engineering[13] and studies of proteins from hyperthermophilic bacteria[14] would imply that higher stabilities would have been achievable by evolution. This observation leads to the conclusion that there is natural selection for limited protein stability[15]. While the precise importance of the various factors on protein folding and stability are still debated, some principles are generally accepted.

Hydrophobic and van der Waals interactions. The dense packing of a protein's hydrophobic core[16] for example, is an important aspect of protein structure, and is presumed to be important for protein stability. Evidence suggests that van der Waals interactions in the densely packed core are of key factors in stabilising proteins[17], and have even been argued to be sufficient for designing well folded proteins[18]. A further force driving the formation of these tightly packed cores is the hydrophobic effect. As hydrophobic amino acid side chains have a favourable transfer free energy from water to hydrophobic solvents[19], it is energetically beneficial for these to group together under the exclusion of the polar solvent. The fact that many proteins tend to denature when exposed to organic solvents such as ethanol underlines the importance of the formation of the hydrophobic core as a key determinant in protein structure stability[20].

Hydrogen bonding. Hydrogen bonds are argued to be one of the most important interactions in biology and chemistry[21]. The strength of hydrogen bonds in proteins is usually estimated to be between 1-4 kcal/mol[22,23]. However, it is argued that hydrogen bonds contribute little to the overall stability of a protein, as residues in the unfolded state are able to form hydrogen bonds with the water and therefore incur an energetic penalty when moved into the hydrophobic interior of the protein. Studies show that the desolvation energy of the backbone hydrogen bonding group is significant[24] and that fulfilling the hydrogen bonding potential by secondary structure formation is necessary to eliminate this penalty rather than contributing positively to the stability. Theoretical studies suggest that the driving forces behind the folding of peptides into secondary structures are the hydrophobic effect and van der Waals interactions[25,26].

Electrostatics. Although salt-bridges appear to be conserved throughout evolution, particularly in solvent inaccessible regions[27], the importance of the interaction of point charges in proteins is contentious. While the contribution of electrostatics to destabilisation due to charge-charge repulsion is well appreciated[28], its effect on the overall stability of proteins is unclear. It is argued that for most proteins the strength of salt bridges is not sufficient to counter the cost of desolvation[29] involved in burying a polar or charged group in the hydrophobic core of a protein. Indeed, it has been shown experimentally in protein engineering experiments that the replacement of charged interactions by hydrophobic ones can increase protein stability[30]. The role of hydrogen bonds appears to lie in defining contact specificity during protein folding[28], rather than contributing to the overall stability. They are, however, as indicated above, inferred to be of importance in proteins from thermophiles[31]. It appears that while single charge-charge interactions are insufficient to compensate for the removal of a charged group from water, a network of interactions, created by optimally placing charged residues, can help ameliorate the detrimental effects of desolvation[32].

Disulphide bridges. As covalent interactions, disulphide bonds potentially contribute significantly to the stability of some proteins. Experiments on ribonuclease T1 show a great loss in stability upon disulphide bridge breakage[33]. However the authors note that this is attributable mainly to the increased entropic freedom of the unfolded stated gained through the lost disulphide bridge, an observation supported by the studies of hen egg white lysozyme[34]. Arguably however, these proteins are not necessarily good representatives of all proteins; they are exceptional proteins involved in the defence against pathogens and as such have very different functional constraints and requirements compared to other proteins.

Loss of function as a cause of disease

Proteins are central and essential to the function of cells, tissues and ultimately the organism. Each one performs a function for which it has been selected by evolution, and a loss of this function is generally regarded as a deleterious phenotype. Mutations can have an impact on the function of a protein, which in turn can cause deficiencies leading to a diseased phenotype. Diseases due to loss of protein function can be metabolic, as in the case of phenylketonuria[35], due to impaired transport through ion channel defects in the case of cystic fibrosis[36] or of regulatory nature, as in the case mutations to the protein p53 in cancer[37]. Critical function loss usually has such a deleterious effect that it will cause death of the organism at a very early point in its development. Non-lethal, but nevertheless harmful mutations cause what can be commonly diagnosed as a disease; such a disease may be lethal at a later point in an organism's life or it may be merely an encumbrance. If a change is not felt by the carrier, it is said to be neutral.

1.1.2 Effects of mutations on proteins

While the process of DNA replication is highly accurate, it is not perfect. With each cycle of DNA replication, a small number of new mutations are introduced which are random with respect to function. Mutations to the DNA in regions which encode for proteins may alter the primary sequence of the resulting peptide and thus have an effect on the biophysical attributes thereof. In the following, the effects of single amino acid substitutions are discussed.

Loss of stability

The loss of stability of proteins is one of the foremost causes of disease. As the proteins are only marginally stable, even small effects on stability alter the thermodynamic equilibrium to make the folded state unstable. Mutational evidence shows that mutations often, if not in the majority of cases, cause major changes to protein stability which are often on the order of magnitude of the absolute stability of the protein[38]. Lowered stability leads to a reduction in a protein's effective concentration, which in turn causes deficiencies in its ability to perform its biochemical function[39]. A prominent example is the deficiency in phenylalanine hydroxylase (PAH), which leads to phenylketonuria. A significant proportion of these mutations are thought to be deleterious due to structure disruption and the consequent loss of stability[40,41]. A further implication of reduced stability is the possible aggregation of unfolded proteins through their promiscuous interaction in the densely populated cytosol[42]. While the cell has safety mechanisms in place to ensure the removal of un- or misfolded proteins, such as ubiquitin mediated degradation by the 26S proteasome, misfolding is commonly associated with numerous diseases[43].

Increased stability

It is known from studies of thermophiles' proteins and from protein engineering experiments, that it is possible to greatly stabilize them. Nevertheless, marginal stability has been favoured by evolution. Thus, it can be presumed that the alteration of the thermodynamic properties of a protein to a more stable state will likely be non-beneficial. Numerous explanations have been proposed to explain this effect.

Increased stability may lead to increased rigidity of the protein which may impinge upon an enzyme's function. The link between flexibility and function is supported by studies on cold-adapted proteins, which display high flexibility and high activity, but only low stability[44]. The phenomenon of induced fit of ligands also relies on flexibility, as for example the hinge motion involved in ligand binding in adenylate kinase[45]. In addition, specific motions, such as subunit and loop rearrangement are essential for catalysis in certain classes of enzymes[46]. Furthermore inhibition, which is often an allosteric effect, is reliant upon rearrangements of proteins. An extreme case is the regulation of pyruvate kinase by means of conformational change[47]. Gains in stability may cause a loss of flexibility which in turn is likely to entail a loss in regulation with obvious detrimental consequences for the cell.

Besides the aspect of enzyme function, increased stability can have an effect on the process of regulation by means of degradation. Many cellular processes are mediated by the presence of key proteins whose removal is an obvious means of regulation. Indeed the cell has developed numerous ways in which to degrade such proteins in a highly regulated fashion, such as for example the ubiquitin system, which degrades a number of proteins known to be of key importance for, amongst other functions, cell cycle control[48]. Increasing the stability of a protein is likely to adversely affect its degradation as the proteasome requires its substrates to be unfolded[49]. Conceivably, an increase in stability could limit the rate at which this process occurs. If the timely degradation of these substrates is altered it can have detrimental effects on cell regulation or adversely affect the dynamic balance between synthesis and degradation in the cell[50].

Misfolding

The problems associated with protein misfolding are intimately linked to those caused by the reduction of stability. Misfolded proteins are in danger of interacting promiscuously with other copies of itself or other proteins in the crowded environment of the cell[51]. Such interactions may lead to the formation of insoluble aggregates in cells or tissues, which can be damaging to the cell or organisms. Highly stable aggregations of misfolded proteins, or fragments thereof, in the form of amyloid plaques are the cause of many conditions; the presence of such plaques in the nervous system is detrimental to neuronal growth and viability, eventually leading to the demise of affected cells or even tissues and can lead to neurodegenerative diseases such as Creutzfeld-Jacob's disease and Alzheimer's disease. Many other localized or systemic diseases are known to be caused by amyloid plaques[52].

Changes in interaction properties

Proteins are often involved in binding other proteins in order to exert their biological function. Naturally multimeric proteins, such as haemoglobin depend upon their multimeric state for their correct function and their assembly requires the correct association of their subunits. Molecular recognition is a decisive factor in many cellular processes such as cell signalling and protein degradation, whereby proteins must recognize specific substrates. Mutations to interaction sites can interrupt many of these processes. The electrostatic properties of a protein, which can play an important role in defining its binding affinity and specificity[53], can be altered by non-conservative mutations. The shape of an interaction site is also important in many proteins, such as transcription factors or endonucleases which are involved in DNA binding[54]. The extent to which such shape changes affect this binding are well enough understood as to allow the design of endonucleases which are capable of binding DNA motifs[55]. The loss of interaction properties can be seen as an extension to the protein stability problem, as the same forces driving protein tertiary structure formation also determine protein quaternary structure.

Loss of active site residues

Enzymes catalyse numerous reactions in the cell, accelerating them by several orders of magnitude. They have to recognize their substrates to catalyse reactions and may also need to recruit necessary co-factors. This goal is achieved by the specific orientation of key residues in their active sites[56], which are generally highly conserved in enzymes of common origin. Furthermore, non-ligand molecules must be excluded from their active sites in order to maximise catalytic efficiency and to avoid unwanted reactions. The conceptually simplest way for a mutation to affect the phenotype is by altering the active sites in enzymes. Changes to residues involved in any of these functions will, if they are not conservative, have detrimental effects on the enzyme's function.

Alteration of covalent modification sites

Finally, many proteins are dependent upon modification for proper function. Proteins may be modulated or flagged for degradation by the addition of phosphate groups, or need to be glycosylated before attaining their functional state. Mutations may affect such posttranslational modifications and be deleterious. For example, evidence suggests that one of the molecular causes of cystic fibrosis is incomplete glycosylation of CTFR[57], which in turn causes its degradation in the endoplasmic reticulum; the resulting insufficiency in copy numbers of this protein gives rise to the disease. Specific protein motifs are recognized by the enzymes performing these modifications and, when altered, may abolish this recognition and cause deleterious effects.

1.1.3 Relevance of point mutation effect prediction

Single nucleotide polymorphisms and disease

Single nucleotide polymorphisms have been commonly defined as mutations at particular sites in the genome, which occur at a frequency of greater than 1% in the human population. These mutations are thought to occur at a rate of 1 every 290 bp in the genome, amounting to a total of 11 million SNPs in the human population[58]. SNPs have been primarily of interest, as they can be associated with certain medical phenotypes such as predisposition to certain diseases[59]. SNPs themselves may, but do not necessarily need to be, causative of disease. Linkage to disease causing variations makes them interesting in the field of medicine as they act as easily identifiable markers which can aid in diagnostics. A certain subset of SNPs however is directly causative of the detrimental effects on an individual's health. In order to identify disease causing mutations for further analysis, it is thus desirable to have a method of predicting their likely molecular phenotype and thus allowing neutral mutations to be sorted from deleterious ones. In addition to existing variation, there is a constant influx of new, spontaneous mutations into the population; humans are estimated to have a high genomic mutation rate, in particular the deleterious mutation rate is estimated to be U=1.6 per genome per generation[60]. The high frequency of mutations and their steady rate of influx into the gene pool underlie the desire to better understand the phenotypic impact of mutations.

Protein engineering

A further use of point mutation prediction is to aid in planning experiments involving mutagenesis. It may be necessary to mutate residues at a particular site in a protein while maintaining its three dimensional structure, or one may want to alter properties of a protein such as stability or enzyme specificity; all of these techniques can benefit from an objective assessment of the effect of point mutations. An concrete application of mutation prediction are experiments involving FRET[61], which uses the efficiency of energy transfer between chromophores attached at particular sites in proteins to determine distances between them. The preferred target for coupling extrinsic chromophores to is the thiol group of cysteines; to conduct an experiment one thus has to remove undesired surface cysteines and introduce them at site of interest[62]. Predicting the impact of mutations can thus help reduce the range of possible mutations to a high confidence subset.

1.2 Modelling the effects of mutations

1.2.1 Impact on protein structure

The extent to which a structure is changed upon mutation is as yet unclear. It appears obvious that rearrangements of the protein are necessary to accommodate changes in amino acid size, either to ameliorate the effects of over-packing or to compensate for the introduction of voids. The exact extent to which this occurs is unclear, but it is desirable to have accurate models for gauging the effect of mutations on the molecular phenotype. The extent to which deleterious mutations change the structure is of interest in rationalising their effect on protein stability. One difficulty in estimating these effects lies with the ability to account for uncertainty in the exact atomic positions of atoms in proteins structures in a meaningful way.

Random variation in protein structure determination

Information on the structural impact of mutations is derived from proteins structures built on the basis of experimental data. The two most frequent sources of data are in the form of electron density maps derived by protein crystallographic means and spatial restraints derived from NMR experiments. Here, the focus is placed on the significance of crystal structures, as the data is far more abundant and the data contains estimates as to their reliability. In protein crystallography, several parameters are provided allowing users to assess the quality of the data and the reliability of the model built. The resolution contains information pertaining to the upper limit of resolvability of the electron density used for building models and it limits the precision with which atoms can be positioned. The R value reflects how the experimental data used to build a model correlates with the data one would expect to observe according to the model. The R-free value is an independent estimate of the R value, whereby a proportion of the data is not used to build the model, but set aside to be used as an independent or free validation set for calculating R[63]. The crystallographic temperature factor[64], or B-factor, reflects the local movement of an atom in the structures; as protein crystallographic data is essentially a time averaged picture of a whole ensemble of molecules, mobile elements will have a wider distribution of their electron density. To fit the data to this distribution, the temperature factor, which reflects the average displacement of an atom, is used to improve the model fit. The temperature factor may also contain uncertainty due to lattice disorder[64]; the copies of a protein in the crystal lattice may not be perfectly aligned and so the electron density reflects the average of this ensemble. Ideally, these factors need to be accounted for when assessing the structural influence of a mutation.

Current understanding of mutational impact on structure

The understanding of the impact of mutations on structures is largely based on case studies on well characterised proteins such as T4 lysozyme. Systematic studies on large sets of proteins, which could lead to a better understanding and culminate in predictive models, are less abundant. Such large scale analyses of the impact of mutations on protein structures are performed by using single point mutant pairs in the PDB[65]. Criterions for assessing the change induced usually structural metrics, such as RMSD or chi angle accuracy, but others have focused on the prediction of other metrics such as changes in protein stability.

Possibly the most notable study of the single point mutations on structure is an investigation on side chain rearrangements in the vicinity of point mutations[66]. Using a sizeable set of single point mutations[67], it was found that up to 95% of mutation sites undergo 2 or fewer side chain rearrangements, many of which were due to the inherent flexibility of the side chains, as determined by observing the variability in a control set of identical proteins.

Predictive methods have been used to model side chain conformation changes in the vicinity of mutations. Feyfant and colleagues[68] modelled side chain rearrangements in the vicinity of mutation sites. When investigating the error they found no dependence of the error on the B-factor. This is presumably in part due to the inclusion of poorly resolved structures with resolutions as low as 3.0 Angstrom, without accounting for structural variation this uncertainty entails. A further limitation on the prediction of error is due to the use of normalized B-factors rather than the original values. Temperature factors are often normalized by subtracting the mean temperature factor over the entire model and dividing by the standard deviation. However, this procedure is not applicable to the problem of comparing structural error, as the temperature factor is an absolute value reflecting uncertainty at a site. Normalisation removes essential information which helps predict the expected random variation.

Bordner and Abagyan use a large dataset of single point mutants for which data was available on the change in stability induced by the mutation[69]. They developed an elaborate method to predict the geometric and energetic impact of mutations. Energetic contributions could be predicted with an accuracy of 1.1 kcal/mol and a correlation coefficient of 0.82 using self-consistency validation. The method has limited applicability for extreme mutations such as the replacement of small residues with large ones. Errors in model building were large for residues with high temperature factors, despite the stringent criteria of removing residues with high temperature factors.

By contrast, Serrano ad co-workers[38] use a very conservative mutation protocol involving as few rearrangements as possible. Using structure models derived in this fashion, the change in stability upon mutation could be predicted with very high accuracy (correlation coefficient between observed and predicted stability changes of 0.83). This suggests that only minimal alteration of the mutation site is necessary for attaining high accuracy in stability change prediction.

The significance of structural variation

As the quality and therefore reliability of PDB structures varies considerably, limits are placed for the inclusion of structures and residues used in these studies. Stratified analyses are performed in order to control for the remaining variation[66]. For example, “binning” residues by temperature factor allows differences between residues with high and low temperature factor to be accounted for, but converting a continuous value into a discrete ones reduces the power of a model. Furthermore, binning by value is generally performed separately for individual descriptors.

To the author's knowledge, the only attempt at controlling error while treating variables as continuous ones was made by Bott and co-workers[70,71]. Using linear regression, they derived a model for correlating the observed variability between equivalent atoms in pairs of structures with the average temperature factor of the two atoms. Using the predicted values for a given average temperature factor and the standard error of the model, the proportion of residues fluctuating due to chance can be calculated. For proteins displaying significant movement, an excess of variability can be observed. Nevertheless, other factors contributing to the random component of the variability are neglected in this model, so there is clearly scope for improvement in this respect.

1.2.2 Impact on protein function

The field of mutation phenotype prediction has been a field of much research in recent years. The aim of many early studies was to interpret and rationalize the effects of single point mutants on protein function. Later methods aimed at modelling the effects on a large scale in a predictive fashion.

Information sources

Modelling or predicting the effect of mutations on proteins requires information about the properties of the site in the molecule at which they occur. Two main sources of information are generally employed, evolutionary data and descriptors derived from protein structures.

Evolutionary information

A multiple sequence alignment of homologous proteins contains a vast amount of information about the evolutionary pressures acting on the proteins over long periods of time. It can be likened to a mutagenesis experiment spanning millions of years. The Neutral Theory of Evolution[72] posits that the majority of substitutions throughout evolution are due to random drift. While recent evidence suggests that a substantial number of substitutions have been driven by positive selection[73], it is generally accepted that the majority will have small effects. Sites which are not of key importance thus diverge over time by means of small, tolerable changes. Functionally and structurally important sites however, cannot tolerate mutations without incurring a selective penalty. The removal of these mutants from a population by purifying selection leads to the removal of variation at such sites. As a consequence, one can infer much information about the likely impact of mutations from the observed variation at equivalent sites in homologous proteins. Tapping into this vast resource requires not only the collection and alignment of appropriate sequences, but also their correct interpretation. In the following, the methods commonly used for predicting mutational tolerance are discussed, along with their potential drawbacks.

Substitution scoring matrices

Scoring matrices reflect the evolutionary propensity of an amino acid to mutate to another. The higher the probability is that two sequences can be inter-converted by a series of mutations, the higher their similarity is. Scoring matrices were developed for the purpose of guiding sequence alignment, whereby those portions of the protein which are most similar are aligned. Scoring matrices were originally derived using the amino acid differences from very closely related proteins[74]. More recent methods such as those used to derive the now commonly used BLOSUM matrix[75], use the frequency of interconversion between residues based on pair frequency counts in blocks of aligned protein segments. These propensities lend themselves to the task of predicting the phenotype of SNPs[76,77]; as deleterious mutations are likely to be removed by negative selection, they will be observed only infrequently, compared to selectively neutral ones. As a consequence, deleterious mutations will score poorly according to substitution matrices, whereas the permissible ones will obtain high scores. The applicability of scoring matrices however is limited and inferior to those scores based on sequence alignments of homologous to the protein of interest[76] (see section *"Position specific scoring matrices"*).

Sequence conservation

A high degree of conservation at a site in a sequence alignment of homologous proteins can be used to infer functional importance, whereas variability is indicative of a lack of functional restraint. Conservation is commonly used in the interpretation or prediction of deleterious mutations[78]. Two caveats must be borne in mind when using sequence conservation. Firstly, the presence of non-orthologous homologues, i.e. proteins with a common evolutionary origin in other species, but which are divergent in function, will falsely imply variability at divergent sites. Functional diversification requires the alteration of key functional sites in the protein, and these are driven to fixation by positive selection. Such sites will have a high degree of variation despite their functional importance. Unfortunately, no current sequence search tool can reliably distinguish between orthologous and non-orthologous homologues. Secondly, highly similar sequences cause sequence alignments to contain redundant information. Ideally, enough time has to have elapsed since the divergence of orthologous protein, in order for the alignment to contain information about neutral sites. An over-representation of recently diverged proteins in an alignment is thus undesirable as it leads to an underestimation of mutational tolerability at neutral sites. Over-representation of sequences is typically dealt with by weighting the information a sequence contains according to its similarity to other proteins in an alignment[79].

Position specific scoring matrices

Besides general site conservation, more specific information can be extracted from a sequence alignment. The abundance of particular amino acids at a given site in a sequence alignment indicates how well this residue is likely to be tolerated. For active sites, conservative mutations may be tolerable if they maintain the interaction properties of the site. Sites in the core, or at the interfaces in the quaternary structure of a protein, may undergo divergence if interactions are preserved and they do not lead to over-packing. Thus, in an alignment of orthologous protein structures, the relative propensity for an amino acid to occur at a site can be derived from the residue profile at its position in the alignment.

Several methods have been proposed to quantify the impact of mutations on the phenotype. The most frequently used is the programme SIFT[80,], which builds alignments while conserving detectable sequence motifs and derives amino acid propensities for sites based on their frequency in the alignment. These scores can be used to great effect in predicting the likely impact of mutations on molecular phenotype[77,81,82]. Similarly, in the research by Bork and co-workers[83], the log likelihood of an amino acid occurring at a site in an evolutionary sequence profile relative to its overall frequency, as implemented in the profile analysis tool PSIC[84], is used to interpret the effect of SNPs. Others have simply used the position specific scoring matrices produced by PSI-BLAST[85] to model mutation impact[76].

Similar caveats must be issued for position specific scoring matrix methods as for conservation scores. Adaptive divergence at a site can falsely imply tolerability, as non-orthologous proteins may differ at key residues, while otherwise being highly similar. There may also not have been enough time for random drift to give rise to variability at neutral sites, leading to an underestimate of variability. A further difficulty is the effect of concerted mutations[86]. If the deleterious effect of a mutation has been compensated for during the course of evolution then the information at both the original site and the compensatory mutation site in a sequence profile can misleadingly imply mutational tolerance to these mutations.

Specific sequence knowledge.

Evolutionary information can be obtained from multiple sequence alignments of homologous proteins. While this method can be fully automated, human intervention may be able to improve their reliability. Many resources are provided to supply detailed, human reviewed consensus sequences, which have been cross-checked against scientific literature in order to validate their functional importance. Such information comes either in the form of regular expressions capturing short protein motifs[87] or longer, family based models which are used to capture domain information, which include the PFAM database[88]. The sequence databases commonly used to gather evolutionary data are continually growing and providing new information, which may not be reflected in such static profiles. The use of dynamically generated sequence alignments therefore potentially has a considerable advantage.

Further detailed annotations, as provided by Swissprot[89] indicate the importance of residues in ligand binding, catalytic function and metal ion chelation. Such information has been incorporated into many mutation annotation schemes[83], but is often omitted in favour of generalized, fully automated information collection methods. The use of specific annotation has some drawbacks making them unattractive. The information provided is limited to proteins which have been experimentally characterised. It thus cannot be incorporated into generalised prediction schemes which lay claim to comprehensiveness and completeness. Furthermore, they reflect only the available information at the time at which these resources are devised. Nevertheless these knowledge resources are valuable for cases in which this information is available and can be evaluated by a researcher.

Structural information

The use of protein structures can be useful in estimating the effects on the biophysics of a protein. Structures provide information about the location of a residue and its environment in a protein. This information can be used to assess likely biophysical implications of the changes which in turn can be used in predicting deleterious consequences.

Deriving structural information

Structural information is usually derived by mapping structures directly to structures from the PDBs[83]. However, the number of experimental structures is approximately two orders of magnitude lower than the number of known sequences, thus there is a deficit in structural information[90]. Information can be derived from homologous proteins[83,91] or, as has been the case more recently, by using homology modelling to predict structures[92]. These approaches typically rely on alignments that are straightforward enough to be found by basic tools such as BLAST[93] or PSI-BLAST[85]. Current state of the art tools use more advanced template finding methods that can detect more remote homologues or increase structural coverage of the target sequence. The bi-annual CASP experiment[94] evaluates to what extent this is the case. Using a double blind set up, predictions of soon to be determined or published structures are made and evaluated by independent assessors using objective criteria. These assessments provide guidelines for choosing appropriate methods for building models for particular tasks.

Protein structures can be used in a number of different ways for interpreting or predicting the effect of mutations. The various uses are outlined in the following section.

Structural rules

The effect of mutations can be rationalized and this knowledge can be used in the prediction of their phenotypic effects of mutations. Many structural rules have been proposed[83,95,96], based on the rationalisation of mutant effects in vitro, for the purpose of predicting or rationalizing the effect of mutations[97,98]. The rules are generally oriented towards detecting the loss of stability, as this is presumed to be the main causative factor of disease[92].

Structural rules employ strict but arbitrary rules for evaluating the change induced by the mutation, and derive a binary assessment of the likely effect. In the following an extensive, but not necessarily complete, list of rules used by the main studies in the field has been compiled. A mutation is deemed to affect structure if one or more of the following are observed:

1. Disruption of the hydrophobic core of a protein by replacing a small chain side by a large one, thus causing over-packing.

2. Cavity formation in the hydrophobic core by replacing a larger side chain for a smaller one.

3. The introduction of a charged or polar residue into the core of a protein.

4. Charge repulsion by introducing an opposite charge.
5. The removal of a cysteine involved in the formation of a disulphide bridge.
6. The replacement of a residue involved in the formation of a salt bridge or polar interactions by a residue not able to maintain the interaction
7. The removal of a hydrogen bonding partner.
8. Introduction a proline into an alpha helix or a site with restricted backbone angles.
9. Change in solubility by the replacement of charged/polar residues at exposed sites with hydrophobic ones.

Furthermore, specific information is used for further inferring the mutational impact of mutations. These include:

10. Mutations at ligand binding, catalytic or allosteric sites in the protein.
11. Alteration of sites of post-translational modification.
12. These rules are used, either alone[96] or in conjunction with evolutionary data[83].

Structural propensities

While structural rules can be used for the purpose of SNP prediction, they rely on broad generalisations and use essentially arbitrary value cut-offs in deciding the likely effect of mutations. A more rigorous approach would be to designate cut-offs based on observed propensities of mutants using known phenotypes. Differences in the site propensities between disease causing and neutral mutations can be elucidated and incorporated into predictive models.

Amongst the first applications of this idea is a study which derived a probabilistic model for predicting SNP effects by using the temperature factor for inferring structural flexibility and therefore mutation tolerability[99]. A further study by Cooper and co-workers[100] used a variety of descriptors to analyse the difference in the biophysical properties of a site. They found that solvent accessibility is the most informative factor for the task predicting a likely clinical phenotype. Other more specialised descriptors such as the site energy and physical strain introduced by a mutation were only of use in a limited number of cases. De la Cruz and co-workers[101,102,103] also analysed sequence and structural propensities of mutation sites; they argued that while structural propensities can be used in rationalising mutation effects, they add little to the information derived from sequences.

Energy functions and statistical potentials

While rules and propensities can be used in predicting SNP phenotypes, they are not universally applicable, are subject to interpretation (such as the exact definition of a hydrogen bond) and do not yield a quantitative assessment of mutations, but rather a qualitative one. Furthermore, they are only capable of binary classification based on two states, stability reducing or neutral and are thus not capable of identifying an increase in stability. The use of objective functions for estimating energy changes is an attractive alternative to structural rules. Two general kinds of functions are available for this purpose. Firstly, empirical force fields tuned using known mutational data are powerful tools which can be used for the task of predicting energy changes. Secondly, potentials of mean force use the Boltzmann relation to infer interaction energies based on contact counts from experimental protein structures.

Empirical force fields such as CHARMM[104], GROMOS[105] and AMBER[106] were devised to perform atomistic simulations of chemical phenomena using an approximation of a molecular system based on Newtonian physics. These force fields incorporate terms with which the forces acting on a particle of a system can be calculated. Such simulation packages can be used to estimate the relative energy change caused by a mutation in a protein[78,107,108], albeit at great computational cost, which is prohibitive to routinely applying it to a large number of mutations. Certain conceptual short-cuts can be taken combining physical force fields with machine learning approaches. By using a physical description of the system and weighting the individual terms in the model using known mutation data, good compromise between physical accuracy and the strength of knowledge based methods can be achieved. This method has been used to great effect to predict the stability changes induced by mutations[38], achieving a correlation coefficient of 0.83 on a database of 1030 mutations.

An alternative approach is to use statistical potentials derived from known protein structures. Using the contact counts between atoms from known structures, the Boltzmann relation can be applied to derive a distance dependent pseudo-energy of interaction for an atom pair[109]. Zhou and Zhou[110] have used this principle for predicting stability changes upon mutation, achieving a correlation of 0.67 between observed and predicted values. Such potentials can be assigned on a per-residue basis, rather than an atomic basis, while still being highly predictive[111]. The use of statistics from the PDB can be applied to other measures such as torsion angles which have been used in conjunction with other terms to predict mutation stability changes[112].

Classification methods

In order to use the information outlined above for classification, the data have to be combined into a model which allows the prediction of the mutant phenotype. The methods used for this task are outlined in the following section.

Rule based methods

The simplest method of combining data into a predictive scheme is using empirical rules to predict the effect of a mutation. For each descriptor used to characterise a mutation, a rule is defined based either on cut-off values for continuous variables, such as solvent accessibility, or on the binary value for categorical descriptors, such involvement in hydrogen bonding. These rules are chosen either on the basis of expert opinion or in order to minimise the error rate for predicting a phenotype using a set of known mutations. Such rules have been used to predict the effect of SNPs based on structural data[96]. While rules represent a very simple way of combining data, they face some drawbacks; rules are often empirical rather than being optimised for the problem at hand and they cannot learn interactions between descriptors unless these are taken into account during rule formulation.

Decision trees

A conceptually simple machine learning tool for combining SNP data into a predictive model is their incorporation into a tree-based decision model[113]. A decision tree repeatedly splits data based on the available descriptors, so as to optimally separate the classes, in this case sorting deleterious from neutral mutations. In graph theoretical terms, a decision tree consists of nodes, at which splits are performed, connected by edges, along which the data subsets obtained from the splits are passed. The resulting model resembles a tree: at its base, or root node, all the data is present; after each split, each subset is passed along an edge, or branch, to the next nodes, where the process is repeated, until a satisfactory level of class separation has been achieved. In the case of mutation phenotype prediction, this separation may be based on descriptors such as sequence conservation or residue burial, which reflect the likely mutability of a site; descriptors which are continuous (e.g. solvent accessibility) are converted to binary values using a decision cut-off value. The rules are chosen by training an appropriate algorithm[114] on known data, which minimises classification error over the training set. Decision trees have a great benefit, in that the structure of the tree, and thus the decision making process, can be intuitively understood by a human. This conceptually simple - as compared to other machine learning tools - method has been used for the mutation classification problem[91,115].

Random Forest

The machine learning tool Random Forest[116] uses an ensemble of decision trees to perform classification and regression. A large number of decision trees are trained on a bootstrap sample of the data. Each tree is grown by subsequently adding nodes to the tree; for each node a subset of the data are used in training and a small proportion of the available descriptors are chosen randomly for decision making. The resulting trees are used to classify new data, whereby the class chosen by the majority of trees in the forest is accepted. Random Forests are not prone to over-fitting[116] thus making it a robust tool for classification. For the task of SNP prediction, Random Forests have been used to combine sequence and structural information[117], as well as geometry descriptors derived from structures[118].

Support vector machines

Support vector machines[119] are a further class of powerful machine learning tools which can be used for classification and regression. This method finds the optimal dividing plane between two classes of points in a sample, by combining the descriptors associated with these points. By the employing kernel functions, non-linear behaviour can be achieved which has been shown to improve the flexibility and accuracy of such models. Support vector machines have been used extensively for the purpose of SNP classification[92,113,117,120,121].

1.3 Problem definition

Modelling the structural impact of mutations

Potentially, much information pertaining to the phenotype of a mutation can be derived from understanding its effect on protein structure. Insights into the extent of rearrangement in a protein allow effects to be interpreted and rationalized. As outlined above, the use of force fields can allow stability changes to be predicted. The accuracy requirements for such models are not clear; while minimalistic approaches to in-silico mutation have yielded very good results in prediction stability changes upon mutation[38], others have used much more elaborate modelling protocols[69].

The interpretation of the effects of mutations on structures is prone to problems due to uncertainty in structure determination by X-ray crystallography. To correctly interpret to what extent proteins undergo rearrangements upon mutation, many factors pertaining to the experimental quality of the data and the expected flexibility of the molecule have to be considered. In the first part of this thesis, a model for predicting the expected level of structural variation is derived. Using alternative structures, i.e. from proteins for which the structures have been solved multiple times, the basal level of random variation is to be determined. This expected level of random variation is employed to isolate and objectively quantify the effects of single point mutations on proteins.

Extending structural coverage through comparative modelling

The number of sequences in the current databases exceeds the number of structures by about two orders of magnitude[90]. As the structure of proteins is more conserved than the sequence, it is possible to use experimental structure of homologous proteins in order to build protein structure models for a vast number of sequences. This technique, termed homology or comparative modelling, is currently the most reliable method in predicting structures. The CASP experiment[122] is an objective evaluation of the current methods by the protein structure prediction community. Given only the amino acid sequence of a soon-to-be-resolved protein, blind predictions of the three dimensional structure are made by the participants.

For structural models to be useful for interpreting biological phenomena, structural accuracy is desired. All participating protein structure prediction methods are thus assessed using standard, objective criteria in order to identify their strengths and weaknesses. For the purpose of rationalising or predicting SNP phenotypes using structural data, homology models need to be built on a large scale, which in turn requires automation. Emphasis is thus placed on assessing automated methods, with a view to using them for the purpose of annotating SNPs. In addition to comparing prediction methods amongst one another, they are compared to one of the standard tool for template identification in homology modelling, PSI-BLAST[85]. The level of improvement of the current generation of fully automated modelling methods over this traditional tool is examined.

Approaches to improving phenotype prediction

The human variant set provided by Swiss-Prot[123] constitutes a vast source of annotated mutation data. An automated method for classifying these mutations would be of great interest as it could be employed to aid in analysing new mutations, as well as being of use in choosing mutations in an experimental setting. While structural data have been used for this purpose, they have been argued to provide little information beyond that which can be derived from sequence alone[101]. The third major part of this thesis is to investigate the use of sequence and structural data in classifying mutations and to derive a robust mutation phenotype predictor.

A new structural descriptor is introduced, namely the predicted energy change upon mutation. This is calculated using a mean force potential applied to protein structures of the wild-type and mutant structure. As mutant structures are only available in rare cases, protocols for modelling their structures are investigated, in terms of which is most informative in classifying mutations.

The use of sequence and structure data to classify mutations is examined using rigorous accuracy tests, aimed at determining their robustness. Training set size dependence of the classifiers is investigated as well as their robustness on unseen data. The top classifiers created here are compared to other methods described in the literature.

2. Results and Discussion

2.1. Assessing the structural impact of mutations

Manuscript in preparation:

Prediction of the random component of variability in protein X-ray structures and its implication for interpreting the effects of single point mutations

Abstract

Discerning biologically relevant differences from background variability is a central issue in biology and underlies our ability to correctly interpret biochemical observations. Here, we analyse the random component of protein structure variation by means of a large scale analysis of alternative crystal structures of proteins. Two goals are pursued in this study. (1) The dependence of local and global similarity of protein pairs on molecular and experimental parameters is investigated, so as to derive a statistical model quantifying the random component of the observed variability. (2) This model is used to investigate the statistical significance of local structural changes observed in pairs of protein structures with single point mutations.

We estimate that approximately 54% of the global variation between alternative structure pairs can be explained by using the crystallographic experimental parameters. Almost 52% of structural variability at the residue level can be explained by a linear regression model incorporating experimental parameters and geometrical structure features. For local, i.e. residue level variability, the crystallographic temperature factor is the main determinant in estimating the expected random variation. Using this regression model for predicting the expected level of variability and thus reducing the level of noise in the comparison of two structures, we show that the effects of conservative versus non-conservative single point mutants on protein structure can be observed significantly better than with uncorrected scores, and that evolutionarily unfavourable mutations cause greater structural deviation than favourable ones.

Introduction

Interpretation of structural movement. Discerning biologically relevant change from background variability is a central issue in biology and underlies our ability to correctly interpret biochemical observations, e.g. recognizing biologically significant differences when comparing protein structures. Structural changes occur in a variety of situations such as upon ligand binding or mutation. Information derived from structural studies therefore has implications for our understanding of the biochemistry and biophysics underlying these phenomena. Attaching biological significance to these changes is usually left to the investigator or not considered at all. Numerous sources of potential noise encountered during experimental structure determination can be identified: (1) Differences in experimental conditions, such as crystallization conditions, pH, temperature, space group and packing of the crystal; (2) differences in data collection, crystal quality and resolution; (3) differences in software algorithms and in the detailed aspects of the refinement. With such uncertainty, there is the danger that the differences observed between structures may not be fully supported by the crystallographic data, and their significance may be over-interpreted[1]. In addition, observed variation also includes residue side chain flexibility, especially at exposed sites, as well as loop movements. While such differences may be supported by the data, it may not be clear whether they are biologically insignificant and thus random, or whether they have biological relevance of which we simply have an incomplete understanding.

Descriptors quantifying experimental noise. Numerous quality indicators are provided by the experimenter, which correlate with the quality of the structure model[2] and can therefore be used in a predictive fashion to estimate the expected level of noise in a structure. The resolution reflects the quality of the crystal and can be limited by the diffraction data collection procedure used. The R value indicates how well the structure factors predicted from the atomic model correlate with the observed structure factors. As the R value can be subject to over-fitting during the structure building process, the statistical tool of jack-knifing is used to derive an R value which isn't prone to over-fitting[3]: a small proportion of the data is not used to build the structure model, and used instead to calculate an independent R value, generally termed the R_{FREE}. At the atomic level, the temperature factor reflects the expected deviation of the atom around the position specified in the structure[4]. Ideally, it would reflect only uncertainties due to fast thermal and slow oscillatory motion, but it can be also reflect experimental error due to static disorder, or mosaicity, of the crystal[4]. Furthermore, the temperature factor can be over-interpreted during the refinement of low resolution data[5], potentially leading to over-fitting. Both the uncertainty due to the experimental method, as quantified for example in crystallography by the temperature factor and R value, and the intrinsic flexibility of the protein, for example solvent exposed side chains and flexible loops, are tightly coupled.

Studies to date. A number of studies have investigated the relationship between temperature factors and structural variation. Bott and Frane[6] used a linear regression model to correlate the distance between equivalent atoms with their temperature factors, which assumed the following form:

$$\log(dR) = a + b * B$$

Where dR is the distance between equivalent atoms, B average temperature factor of the atoms, and a, b are the regression coefficients derived by linear regression. Using this approach they could estimate the expected level of variation for a given atom based on its temperature factor and convert the raw distance into a Z-score, indicating the significance of the variation. This method has been used to identify regions of proteins which have higher than average variability, and are thus more likely to be of biological importance[7].

Stroud and Fauman[8] analysed the differences in atomic positions in crystal structures, for which multiple copies of the same peptide are present in the unit cell. They extracted the difference in atomic positions after superposition of the inflexible cores and fit their data to an empirical function of the form:

$$dR = a + e^{b \cdot B}$$

Where again dR is the distance between equivalent atoms, B is the temperature factor of the atoms, and a, b are the regression coefficients. They found a strong correlation between the atomic temperature factors and the observed variation between structures in their set of 18 protein structures. A further observation was that the restraints imposed on atoms by their connectivity atoms are also important in estimating the expected structural variation. In a more recent study, Halle[9] shows that the temperature factor is tightly coupled to the local packing of the protein. Using a non-redundant set of high quality protein structures, Halle derived a model for estimating temperature factors based on the contact density on a per-residue basis. The model is able to capture many intrinsic motions in protein structures and its predictive power illustrates that the packing density contains much information pertaining to the flexibility of structures. Geometrical features capturing packing of a site in the protein structure can potentially contain additional information not contained in the temperature factors.

Implications for single point mutations. It has been observed that the structures of homologous proteins become more similar with decreasing divergence. However, alternative structures of the same proteins are not identical, but still vary to a degree similar to that of closely related proteins[10]. The special case of proteins differing by one difference, i.e. single point mutations, has been investigated only in few systematic studies. These studies employ strict but arbitrary experimental parameter cut-offs, for example placing limits on resolution and R factor, to exclude unreliable structures during dataset selection or use temperature factors cut-offs to exclude unreliably resolved atoms or residues. For example, the structural variation caused by mutations in single point mutant structure pairs has been investigated by quantifying the degree of variation in side-chain movement upon mutation[11]. It can be shown that while single point mutant structure pairs do vary compared to structure pairs of the same sequence, the background variation observed significant. Furthermore, a statistically significant dependence of side chain flexibility on the accessibility of the residue exists, which needs to be accounted for. This study however relies on using only residues with low temperature factors and does not attempt to derive a predictive model for the expected level of variation on a per site basis. In addition some of the measures used are potentially uninformative (chi angles), limiting the broader applicability of these findings. The transformation of continuous variables (chi angles) into binary ones ("correct" or "incorrect" rotamer) may underestimate the variability. Furthermore, their study was limited to proteins with a fixed backbone, which places an artificial limit on the maximum degree of variation which can be observed.

In this studt we extend the analyses and ideas outlined above. Several linear models are investigated, which incorporate an increasing number of descriptors; these include relevant crystallographic data, as well as geometric measures to derive a general model of variability. The applicability of such predictive models is tested by means of a large scale analysis of equivalent proteins. Finally, a generalized predictive model is used to determine the significance of the effect of single point mutations in proteins structures.

Results

Characterizing local variability. Local variability can come in the form of backbone shifts, rotational movement of the side chain and combinations thereof. In previous studies, others have treated these separately or focus on only one aspect such as the side chain angles. Here, we used the root mean square deviation (RMSD) of equivalent residues in the two alternative structures, after a global superposition based on the alpha carbon atoms, as the measure of variation at the amino acid level between alternative structures of proteins. A linear model was used to relate amino acid variability to the various descriptive factors, in order to derive a predictive model for estimating the expected variability and to determine the importance of each predictive factor in doing so. The predictors, as outlined in table 1, fall into two categories: descriptors for structures derived from X ray crystallography, which describe the quality of the data and the confidence in the refinement of the structure, i.e. resolution, working and independent R value (R-work and R-free) as well as the temperature factors of the residues. The second category consists of those which are derived from the geometry of the structure itself and includes solvent accessibility parameters and the number rotatable bonds of the amino acid, which reflect the degree of freedom a residue has for adopting alternate conformations. Multiple linear regression was used to fit the data and predict the expected level of variation in a protein structure.

Dataset creation. To examine the variability in protein structures we created 3 sets of identical proteins and one set of single point mutants. (1) The first set allows us to examine the degree of noise expected between structures in identical conditions by using protein chains from the same PDB entry (Same PDB entry Set, abbreviated to SPS). These proteins will be equal in terms of experimental conditions and quality. Not only are the experimental parameters the same (resolution, R-values), but also the experimental conditions, such as pH, temperature and solvent. This permits the examination of the local variability in absence of differences caused by altered physical conditions of the experiment, as well as the effect of experimenter bias. For this set, only PDB entries were chosen for which no non crystallographic symmetry (NCS) information in the form of restraints or constraints was used during refinement, thus ensuring that the similarity of individual chains was not an artefact of the refinement but a true reflection of the experimental data. (2) To examine the effect of different experimental conditions on the structure of proteins, a second set composed of pairs of alternative structures of same protein taken from different PDB entries was analysed (Different PDB entry Set, abbreviated to DPS). For these protein pairs, the experimental conditions may differ between both members and should allow us to identify the degree of additional variation thereby caused. In addition, the effect of experimental parameters describing the quality of the data and refinement can be examined. It has been observed that structures derived by the same authors differ less from one another than structures solved by independent groups[12]. It thus stands to reason that this effect may affect the predictability of the noise component, which might be expected to be reduced due to unwarranted similarity of the structures. Thus the DPS was chosen to include only structures, in which the authors list has at least one name in common. (3) In contrast to this, a third set of protein pairs was created (Different Author Set, abbreviated to DAS), which consisted of pairs of peptides from different PDB entries which had been solved by different authors. This set was used to determine how independently solved structures vary and how well this variation can be captured by our regression models. The sets of alternative structures were selected and screened by the quality measures outlined in the Materials and Methods section of this report.

RMSD as the measure of similarity. Each pair of peptide chains was superposed using least squares fitting of the common subset of alpha carbon atoms. Based on this superposition, the RMSD values for each residue were calculated and recorded. A lower RMSD cut-off was used to exclude structure pairs with global alpha carbon RMSD values of less than 0.01, because these are so similar that there is good reason to assume two such structures were not solved independently. An upper cut-off of 10 Ångstrom for the global alpha carbon RMSD was used to exclude structures with very large and obvious structural deviations, as the least squares global superposition of two such structures is essentially meaningless and one can safely assume a significant change has taken place.

Local descriptors. The local descriptors, i.e. ones calculated on a per-residue basis, in order to capture the variability of specific sites in the protein (the local descriptors can be found in table 2). The maximum atomic temperature factor for in a given residue was used to capture the expected structural variation due to motion in the crystal. Using the maximum atomic temperature factor was preferable to taking the average over the entire residue, as it provided a better model fit than did the average residue temperature factor (data not shown). The alpha carbon (Cα) density is calculated by counting the number of alpha carbons in a radius of 15 Ångstroms around the alpha carbon of a residue. It is a measure of the general packing density around the residue and thus its ability to adopt alternative conformations. The rotatable bond number reflects the number of side chain dihedral angles which can vary and is a measure of the maximum flexibility of the amino acid. Information on crystal contacts was included in binary form, i.e. the value assumes 1 if a residue is in a contact and 0 if not. Previous studies arrive at conflicting verdicts as to the effect of crystal contacts. While Stroud and Fauman[8] find little evidence for their effect on variability, a more recent study using a larger dataset[12] showed that they can cause large movements. In addition, the steric confinement of amino acids at crystal contacts lowers their temperature factors leading to a further underestimation of the expected variation. Although the impact in absolute terms is comparatively low, the trends observed are highly significant when using large sample sizes. In this study, Contact state 1&2 are binary descriptors indicating whether the equivalent residues in the respective structures are in a crystal contact. The descriptor "Equal contact state" assumes the value 1 if equivalent residues in the respective structures are either both in, or both not in a crystal contact. A residue is deemed to be in a contact if it is closer than 4 Ångstroms to another chain in the unit cell or is in contact with a chain brought into contact by applying the appropriate crystallographic symmetry operators.

Global descriptors. Global descriptors, i.e. ones which apply to all residues in a PDB entry, can give insight into the expected level of noise in the structure in general. The resolution of an X-ray diffraction experiment indicates how good the quality of the data is for deriving the structure, and depends on crystal quality and experimental apparatus used. The values R and R-free values are measures of how well the structure fits the experimental data, whereby the R value is a self-consistency test using the same data for fitting as for testing. The R-free value represents a more statistically sound goodness-of-fit measure as it is derived using a subset of data not used during the refinement process[13]. The mean temperature factor of the structure is an additional indicator for unreliability in the refraction data or for bad refinement.

Regression analysis. The local variability was modelled using log linear regression analysis. The general form of the equations used for predicting local variability is as follows:

$$Log\,(rRMSD\,) = Intercept\; + w_1 * \; factor_1 + w_2 * \; factor_2 + ... + w_N * \; factor_N$$

Hereby, rRMSD is the RMSD between equivalent residues. The factors factor$_1$ to factor$_N$ are the independent descriptors used for predicting variability. The intercept and regression coefficients w$_1$ to w$_N$ are determined by means of least squares fitting during linear regression. For the local descriptors, values were taken from only one of the two protein structures. By including terms from the comparison of both structures, there is the possibility of over-fitting as the combination provides explicit evidence for a structural difference. For example, if flexible portion (such as a loop) of the protein in contact with the rest of the protein in one structure, but is unrestrained in the other, this will be reflected in the difference in the Cα density of equivalent residues in this area, despite the fact both conformations are in principle accessible to both structures. This difference in Cα density will be direct evidence that the two differ in conformation, rather than being an implicit descriptor of uncertainty. For the global descriptors, there are two values for each of the two structures being compared, one of which will be the "worse" or limiting, the other will be the "better" value. During regression, better results can be achieved (data not shown) by treating the "worse" and "better" values as such, rather than treating them as factor 1 and factor 2 depending on the arbitrary order in which the pairs of PDBs are assigned.

Protein level regression model using local features. In order to estimate the random component of the variability, the local regression model was first applied to individual structure pairs. Stroud and Fauman[8] noted that any coefficients derived during regression by machine learning methods vary from protein to protein and may not be transferable, and they therefore make no attempt to derive a general regression model by pooling variation data from different structures. To examine how well the local regression model can be fit to individual protein pairs, regression analysis was performed individually on each of the pairs in the different sets. The resulting protein level R-squared values were recorded and the distributions for the various test sets were plotted in a histogram (figure 1). The distributions are all remarkably similar for all sets; all distributions peak at the same bin (0.45+-0.05) and only the peak in the distribution for the SPS is at a slightly higher bin (0.55+-0.05). To test whether there was a significant difference between the distributions of R-squared values for the various sets an all-against-all comparison was performed by means of a two-sided unpaired Student's t-test. The t-test (see table 4) indicates that none of the distributions of R-squared values differ significantly from one another at the 99% confidence level, and only those of the DAS and DPS differ significantly at the 95% confidence level. A further influence on the protein level correlations could be quality of the global structure fit; large changes may worsen the superposition and decrease the protein level correlations. To test whether there was a link between the global variation and the per-protein regression model quality, the correlation between the logarithmic global Cα RMSD and the protein level R-squared values was calculated. The protein level regression model fit was shown to be correlated with the global RMSD only in the case of the SPS set, for which a statistically significant but generally uninformative correlation was found (R-squared values of 0.041 and p-values of 3.857e-15).

Generalized regression model using local features. To test to what extent the regression parameters are transferable, all residues from each of the pairs in the respective samples were pooled for regression. The resulting regression models (table 2) have lower explanatory value (as measured by the R-squared value) than the best models derived on a per-pair basis. In the smallest sample, the DAS, the regression model could explain 42.7% of the observed variability, which corresponds to a multiple correlation coefficient of 0.65 between observed and predicted variability. By contrast, the distribution of the R-squared values derived on a pair-wise basis has its mode at 0.55 (figure 1); thus the fit of the regression model for the pooled sample is slightly worse than might be expected from the separate pair-wise correlations. For the DPS and SPS, the multiple R-squared values for the generalized models are 0.274 and 0.214 respectively, indicating a further loss of explanatory power with increased sample sizes. Amongst the local descriptors, the local temperature factor was the most decisive in predicting variability, followed by the Cα density and rotatable bond number.

Generalized regression model using local and global features. In the comprehensive regression model, the local descriptors were used together with additional global descriptors, as listed in table 3. In the case of the DAS, a linear model using both local and global descriptors could account for 51.6% of the variation observed, corresponding to a correlation coefficient of 0.718 between observed and predicted values. The better of the crystallographic parameters, R-free, resolution and mean temperature factor, provide more information than the limiting values. For the DPS, a total of 36.7% of the variation in the data could be explained, when including the global descriptors, whereby all descriptors had associated p-values of lower than 2e-16. As for the DAS, the better mean B-factor was more informative than the limiting value for the DPS set; however, the limiting R-free and limiting resolution were more informative for the DPS set than the better value. For the SPS, the R-squared rises to 26.5% from the 21.4% observed for the purely local model. One of the possible reasons for the observed difference in correlation between the sets is a strong difference in the global similarity. As the local variation is linked to the global variation, the observed change in regression model quality could be due to variations in the global RMSD distributions of the different sets. Figure 2 shows the distribution of the global Cα RMSD values for the various sets. An unpaired, two-sided t-test showed that the sample means not differ significantly at the 95% confidence interval (the most significant t-test score is p=0.172). The standard deviation of the logarithmic global Cα RMSD distributions did however vary (0.881, 0.921 and 0.932 for the DAS, DPS and SPS respectively), which would conceivably affect the standard error of the linear models fit to the pooled residue samples.

Predicting global changes. The measure of local variability employed here, the RMSD of equivalent residues, is dependent to no small degree upon global variability. The inclusion of global descriptors may thus be directly predictive of the global RMSD. The dependence of the global RMSD on the global descriptors was investigated by means of log linear regression. The RMSD of alpha carbon atoms from equivalent residues was used rather than the all atom RMSD, so as to exclude side chain movements from the calculation. The following regression formula is used for characterizing the variability:

$$Log\,(cRMSD\,) = Intercept\ + w_1 * \ factor_1 + w_2 * \ factor_2 + ... + w_N * \ factor_N$$

Hereby, cRMSD is the global Cα RMSD. The descriptors $factor_1$ to $factor_N$ are the descriptors listed in the table 5 and the intercept and regression coefficients w_1 to w_N are determined linear regression.

For the various sets, different descriptors were informative as to the expected variability (table 5). For the DAS, the better of the two mean temperature values of the two structures was the only significant factor for predicting global RMSD. The regression model explained 53.6% of data using mainly these descriptors, which means that the expected RMSD can be predicted with a correlation coefficient of 0.73. For the RMSD values in the DPS, the proportion of variation explained dropped dramatically to 19.6%; whereby the limiting R-free value was the most informative in predicting RMSD. In the case of the SPS the resolution was the most predictive feature for estimating the variability of the structures, however it only accounted for a total of 8.7% of the variation. It has been suggested that the RMSD between two proteins may be dependent on their size[14]. To test whether this was the case, the correlation between protein length and logarithmic global Cα RMSD was calculated. Only in the SPS a significant correlation was observed, but the linear model explained only 2% of the data (R-squared=0.02271, p-value=5.214e-09; data not shown for the other two alternative structure sets).

Application of the local RMSD linear regression model to single point mutations. As shown, the random component of the variation between two otherwise equivalent protein structures can be modelled using a linear combination of terms pertaining to the local and global quality of the structure and the expected local flexibility. This statistical model can be now applied to the problem of estimating the level of noise in structure comparisons, by deriving a normalized RMSD value for each pair of residues. The expected level of variation is predicted using the linear model derived from the DAS using the complete set of descriptors (table 3). As outlined in materials and methods, this predicted log RMSD value is subtracted from the observed log RMSD value and the difference is divided by the standard error of the linear regression model in order to derive a Z-score, i.e. the “normalized” RMSD.

This practice was applied to the analysis of the structural effects of single point mutations. A set of structures of point mutation pairs was selected from the PDB, for which all structures had the necessary descriptors to perform the predictions. In order to examine the effect of mutations, a comparison was performed of all sites in the vicinity of a mutation, i.e. those pairs of equivalent residues for which at least one had at least one atom 4.5 Angstrom or closer to any atom in the side chain of the residue at the mutation site. To describe the severity of the mutation, the SIFT[15,16] score was calculated for the amino acid at the mutation site in each of the two variants. The SIFT score is similar in concept to a position specific substitution matrix, in that it provides a measure of evolutionary “goodness-of-fit’, i.e. how well a particular amino acid can be accommodated at a site, according a multiple sequence alignment of homologous proteins. The difference in SIFT score of between the “wild-type” and the “mutant” residue determines whether a single point mutant pair is deemed “conservative” or “non-conservative”. In this study, single point mutant structure pairs with a difference in SIFT score of greater than or equal to 0.8 were designated non-conservative, those with a SIFT score difference below this threshold were deemed to be conservative. The residues occurring in the vicinity of the mutation site were thus assigned to the conservative or non-conservative set. According to this criterion, one would expect the non-conservative mutations to be more disruptive to the structure than the conservative ones.

For the first test of the effects of mutations, all residues in the vicinity of the mutation sites were used. Figure 3 shows the distribution of both the log RMSD as well as the Z-score for the two subsets. The difference in the non-normalized log RMSD shows that few differences are observable between the two different sets, and a Student’s t-test (table 6) indicates that there is no significant difference between the samples (p-value=0.6056). The Z-score by contrast eliminates a portion of the random component of the variation, thus a difference between the two distributions can be observed. The t-test between the Z-score values in the two sets revealed a clearer picture than the simple log RMSD (p-value=$3.275*10^{-16}$). This illustrates that slightly larger changes take place in the non-conservative mutants than in the conservative ones. This effect was only small, and is only apparent upon reducing the random component of the log RMSD values. The effect became more prominent, when only considering the largest residue movements. For each pair of structures, only the pair of mutation site neighbours with the highest displacement according to the Z-score was selected. As shown in figure 4, the difference in the variation between the conservative and the non-conservative sets was more pronounced for the Z-score than for the log RMSD. The t-test (table 5) shows that the difference between the value distribution of the non-conservative and conservative set is insignificant when the log RMSD is used as a metric (p-value=0.9718), but is highly significant for the Z-score (p-value=$2.850*10^{-4}$).

The ability of the local descriptors to correctly predict the random component of the structural variability varies widely for individual protein pairs (figure 1). If the degree of structural variability cannot be predicted, the normalisation might be expected to be less effective and the Z-scores will be uninformative. For the final test we thus excluded pairs in which this is the case, by selecting a subset of the mutants for which the per-protein R-squared values were greater than 0.7. Figure 5 shows that according to the non-normalised log RMSD, the conservative mutations had greater structural variability than the non-conservative mutation ones. Once the normalisation was applied, this trend was reversed and variation around the non-conservative mutation sites became greater than that of the residues in the vicinity of conservative mutations.

The ability of the Z-scores to discern between conservative and non-conservative was contrasted with the performance of a basic model, which merely distinguishes between buried and exposed residues. A linear model using only this information was trained on the DAS and applied to the mutant dataset in the same way as the comprehensive model. The Z-score from this basic binary model lacks much of the power of the comprehensive model. While it can discriminate between non-conservative and conservative mutations when using the full set of residues, the significance of the difference of the means (p-value=0.005) is not as high as when using the comprehensive model for normalisation (p-value=3.275e-16). When using only the residues which rank highest by Z-score, the simple model cannot discern between the Z-score distributions of conservative and non-conservative mutation site residues (p-value=0.1211).

Discussion

Local scores. The dependence of local variation of protein structures on factors describing the global and local quality of experimental structures has been analysed using log-linear regression. Applied on a pair-wise basis, the local model displayed an extreme range of goodness-of-fit values, ranging from multiple R-squared values of 0.785 for the best pair in the SPS set, down to values insignificantly different from zero. This demonstrates that the random component of the proteins structures is not always well predicable from the local descriptors. The temperature factor, which of all descriptors contains the largest proportion of the information pertaining to the variation, does not contain the same information for all structures. The high correlations observed do however show that the approach is valid in principle and that the local descriptors are capable of capturing much of the expected variation. But the high degree of heterogeneity in how the temperature factors are calculated make the general model sub-optimal. This heterogeneity promises to be remedied by refinement of the protein structures in the PDB using the original experimental data[17]. By using a uniform approach to structure solution (using the same protocols and software), bias should be eliminated and consistent estimates of structure quality can be obtained. Conceivably, the experimental values from structures solved in such a manner would improve the fit of the linear regression models proposed here.

As to the applicability of the general regression model, i.e. in which all data points from the respective sets were pooled, we find that using only local information, up to 42.7% of variation can be observed in the smallest data set (DAS). The DAS had a relatively tight distribution of R-squared values in the protein level regression, indicating that the temperature factors in the pairs contain similar information, resulting in a comparatively good fit for the pooled data model. Linear modelling fares less well when applied to the pooled data from the DPS and SPS, which have broader distribution of per-protein correlation coefficients. It is notable that the distributions of the regression coefficients, which are derived on a per-protein basis, did not differ significantly between the sets, indicating that no set is systematically worse in terms of the reliability of their local information. It appears that other factors such as the differences in global experimental parameters vary to such an extent that the pooled regression analysis is unable to attain good correlations.

The inclusion of global factors such as resolution and R-values improved the fit of the general model, so that it could explain over half the local variation observed (51.6%) in the case of the set of structures solved independently by different groups. The predictive power of the model drops when applied to structures solved multiple times by the same group, and even further when applied to chains from the same PDB entry. While it would be tempting to assert that proteins solved by different groups represent truly independently solved structures, in which the temperature factors reflect the uncertainty in structure refinement, and that author bias introduces artefacts which reduce the correlation between random variation and crystallographic and geometric descriptors, the link is more complex. The regression analysis performed on a pair-wise basis indicates that the general behaviour of the random component of structural variation with respect to the local descriptors is not significantly different between the various sets; the t-test result indicates that the distributions of pair-wise correlation coefficients derived from the various sets do not differ significantly at the 99% confidence level and only differ at the 95% confidence level between the DAS and DPS. The difference in the explanatory power of the various general models is thus likely to be due to the larger number of protein pairs included in the sample and that the extreme differences in the per-protein correlation coefficients distribution adversely affects the general model. A further point is made by Stroud and Fauman[8] who remark that regression coefficients derived from one pair of proteins may not be applicable to another. The more diverse a set is, the less likely it is that the parameters derived will be applicable to all structures, and as a consequence the fit of the regression model will be decreased.

A further problem is that the regression model tended to only work well for globular proteins, for which the superposition is driven by a bulky, largely immobile interior.

Superposition methods based on a least squares fitting are largely inappropriate for non-globular proteins, as large movements lead to a meaningless superposition and a large global RMSD. The variation due to the poor superposition can outweigh the differences due to flexibility or ambiguity. Manual curation of the dataset, eliminating such cases might lead to better regression results, but “cherry picking” appropriate data points would in the authors’ opinion make the reasoning in this study circular. One implication of this large spread is that, when assessing differences between structures, it may be necessary to develop protein family specific models for assessing significance rather than relying on a regression model derived from a pooled sample of heterogeneous structure pairs. The fit of the regression models might improved by using an iterative superposition method. This would eliminate the effect of structure pairs, in which one member has undergone a domain shift or hinge movement. These effects are currently unaccounted for except by the imposition of an upper RMSD cut-off; to what extent this procedure would improve the regression models is as yet unclear.

Global variation prediction. The global RMSD between two proteins is correlated with the local RMSD; if a superposition has a low RMSD, the local RMSD of each residue will be correspondingly lower. In order to assess how much of the global component the regression analysis has captured, the correlation between the global RMSD and the global predictors was examined by log-linear regression. The regression analysis for the global RMSD values showed that the global descriptors can explain up to 53.6% of the variability in the RMSD in independently solved structures (DAS) and virtually all information is provided by the better average temperature factor of the two structures being compared. The proportion of the data explained by such a linear regression model varied between the data sets. The variability in the set derived by non-independent authors (DPS) was determined mainly by the limiting R-free value, highlighting the importance of this independent measure of structural goodness. The similarity of chains in the SPS was governed by the resolution of the experiment, but it only explained a small fraction of the variability. The reason for this reduction in explanatory power of the regression model appears to be linked to the increased global variation of the structures in the set. For the alternative structure sets used here, the standard deviations of the logarithmic Cα RMSD distributions grow with increasing sample size, even though the sample means remain the same. The reason why the increased variation is not being explained by the regression model is unclear and may be related to the diversity of the protein type in the samples. The increased standard deviations of the logarithmic Cα RMSD distributions in the larger sets may also have an effect on the local residue variability.

Application of the local variability model to single point mutants. The task of identifying significant differences between structures of proteins with single point mutations is hampered by a considerable degree of noise due to experimental limitations and random motion in solvent accessible parts of the protein. By applying the comprehensive linear regression model for predicting the expected level of variation, a large degree of this variation can be accounted for, consequently isolating the signal due to the actual structure change. The log RMSD was converted to a Z-score by subtracting the predicted variation and dividing by the standard error of the regression model, using the linear model trained on the DAS. Applied to the full set of mutants, the difference in the distribution of Z-scores between the conservative and the non-conservative sets was significantly different, whereas the log RMSD distributions could not be told apart. The effect was exacerbated when regarding only the residue with the highest deflection from each protein in the set. In this case, the differences between conservative and non-conservative mutation sites according to the log RMSD were insignificant, whereas for the Z-score, the significance is high and the two distributions were discernable clearly by eye. A simple linear model using only the information on whether a residue was buried, was only able to distinguish between conservative and non-conservative mutations when the entire set was used, and the significance of the estimate was far lower than for the comprehensive model. This shows that the standard binning procedures usually used are far less powerful an approach, and that factors describing the experimental quality are essential in discerning significant from insignificant change.

As many proteins in the single point mutant set did not boast a strong correlation between the observed variation and that expected according to local descriptors, the noise cannot be predicted accurately for these pairs. To best visually display the power of the normalization procedure, a subset of mutation pairs was chosen, for which the correlation between observed and expected variation for all residues is high. In these cases, the difference between conservative and non-conservative mutations was counter-intuitive, when using the logarithmic RMSD as a criterion; the conservative mutations showed a larger log RMSD than the non-conservative ones. When applying the Z-score, the trend was reversed and the non-conservative mutations showed larger differences in residues close to the mutation site than those in the conservative mutation set. This displays how reliance on RMSD, without accounting for expected variation can be misleading. Accounting for the noise during structure comparison thus allows the noise component in the observed structural change upon mutation to be reduced. The signal from the actual change can be increased and previously undetectable differences can be made detectable. A caveat exists however, in that the noise component between two structures must be well correlated with the local descriptors used in predicting it. In the case of single point mutant structure pairs where this correlation is high, the effects of the normalization were most effective.

From a biological point of view, the observed patterns of structural change imply that even small movements in the structure are likely to incur a penalty in the evolutionary fitness of the variant. Presumably, this is due mainly to changes in stability, as this is assumed to be the main cause of disease phenotypes associated with SNPs in humans[18]. The small differences observed illustrate the intolerance of protein structure to non-conservative mutation.

Summary

The random component of the variability between structure pairs could be predicted by a residue based statistical model using experimental and geometrical descriptors. When regression parameters were derived on a per-protein basis, almost 80% the structural variation could be described in exceptional cases. Pooling data generally reduced the correlation, but the generalized model using the comprehensive set of descriptors could still explain approximately half the variability. Over 50% of the global variability, which is correlated with the local variability, was explained by factors describing experimental quality. The local linear model can be used to derive Z-scores which normalize structural variation. It was shown that when using these Z-scores the mutation site variability is higher on average for evolutionarily unfavourable mutations than for favourable ones. This effect was not apparent using the standard RMSD, as the noise component outweighs the signal. The effect was visually most prominent for structure pairs in which the structural variability is highly predictable using local descriptors.

Materials and methods

Dataset. The dataset for the analysis consists of all chains of all structures of the PDB[19], which consist of complete residues (no structures containing only Cα traces were included) do not contain ambiguously labelled residues, provide R-work and R-free values, as well as the resolution of the experimental data. All proteins had to be able to be assembled into continuous peptides, with the peptide bonds not exceeding 2.0 Angstroms in length. The peptide sequence had to match or at least be present in the corresponding the SEQRES entry in the PDB file. For pair-wise comparisons, the structures were reduced to the common subset of residues in the analysis in these cases. Identical sequences were grouped and 2 representatives were chosen at random from this group. The PDB entries were required to provide the appropriate crystallographic symmetry information for calculating crystal contacts. Any ligand with more than 5 atoms had to be present in equal numbers in both structures, without imposing RMSD limits on the ligands. We further filtered the set to include only PDBs which contained copies of the same peptide in their unit cell, i.e. all peptide chains in the PDB had to have the same SEQRES entry; in order to exclude peptides which were part of heterocomplexes (i.e. bound to nucleic acids or non-self proteins). Proteins in different complex states can undergo large conformational changes and their intrinsic properties at interaction sites may no longer describe random variation, but biologically significant changes induced by the formation of the hetero complex. The datasets were compiled from this set of filtered chain pairs. The first set (the “Same PDB entry Set”, SPS) consisted of pairs of chains taken from PDB entries which contained two or more copies of the peptide, and for which no non-crystallographic symmetry restraints were used during refinement. The second set (“Different PDB Set” or DPS) was composed of pairs of different PDB files, but which were not solved by independent authors. As the DAS could in principle be a subset of the DPS; only pairs of structures were chosen for the DPS, which shared at least one author. The third set contained only chains taken from PDB files which had no author in common according to the AUTH entry of the PDB file. For each protein sequence in these sets, two representative structures were chosen at random. In order to avoid overrepresentation of certain protein families in the sample, each set was culled at the 25% sequence identity level using the PISCES server, provided by the Dunbrack group[20,21]. The single point mutation set (SMS) was created using the same filters as for the other sets, except that they were not culled at the 25% sequence identity level; instead the structures were chosen, so as to contain only unique mutations; for each mutation occurring in a particular sequence, only one pair of representative structures were chosen, thus avoiding the inclusion of pairs structures comprised of several alternative wild type structures paired to the same mutant.

Structural descriptors. Temperature factors were extracted from the atomic entries in the PDB files. The maximum atomic temperature factor per residue was used for statistical modelling. Temperature factors were capped at 150.0. The alpha carbon density of a residue is defined as the number of alpha carbon atoms in a 15 Ångstrom radius of the residue's alpha carbon atom. For handling PDB files and extracting structural information, the Bio.PDB module[22] of the Biopython suite was used. The program 'contact' from the CCP4 suite[23] was used to calculate crystal contacts. Solvent accessibility was calculated using the program NACCESS[24]. Residues were classed as buried if their relative solvent accessibility was below 5%, the remainder were classed as exposed.

Superposition and root mean squared deviation. Structure superposition was performed using the SVDSuperimposer class provided by the Bio.PDB python module, which finds the least squares fit of two sets of points by singular value decomposition. Global superposition was performed using all alpha carbon atoms in common between two structures. The root mean squared deviation (RMSD) between two sets of atoms is defined as:

$$RMSD = \sqrt{\frac{\sum_{i=1}^{N}(x1_i - x2_i)^2 + (y1_i - y2_i)^2 + (z1_i - z2_i)^2}{N}}$$

Where N is the number of atoms used, and the x, y, z values are the three dimensional Cartesian coordinates of corresponding atoms in the two structures. When calculating the

RMSD for a pair of residues, the ambiguity due to UIPAC nomenclature (Arginine NH1/NH2 etc.) was accounted for by assigning equivalent atoms so as to minimize the residue RMSD.

The logarithmic residue RMSD values were converted to Z-scores as follows:

$$Z\_score = \frac{RMSD_{observed} - RMSD_{predicted}}{\sigma}$$

Where the predicted RMSD is calculated from the complete set of descriptors using the linear model derived using the DAS, and σ is the model's standard error. All Statistical calculations and modelling were performed using version 2.6.2 of the R package[25].

Residue preference scores. SIFT[15] scores were derived using the Uniref90 database[26], a non-redundant database which is culled at the 90% sequence identity level. If the difference in the SIFT scores of the residues was less than 0.8, the mutation was classed as conservative, otherwise it was classed as non-conservative.

References

1. Kleywegt, G.J. Validation of protein crystal structures. Acta crystallographica. Section D, Biological crystallography **56**, 249-65 (2000).
2. Brown, E.N. & Ramaswamy, S. Quality of protein crystal structures. Acta crystallographica. Section D, Biological crystallography **63**, 941-50 (2007).
3. Brunger, Axel T The free R value: a novel statistical quantity for assessing the accuracy of crystal structures. Nature **355**, 472-4 (1992).
4. Ringe, D. & Petsko, G.A. Study of protein dynamics by X-ray diffraction. Methods in enzymology **131**, 389-433 (1986).
5. Kleywegt, G.J. & Jones, T.A. Where freedom is given, liberties are taken. Structure (London, England : 1993) **3**, 535-40 (1995).
6. Bott, R. & Frane, J. Incorporation of crystallographic temperature factors in the statistical analysis of protein tertiary structures. Protein Engineering **3**, 649-57 (1990).
7. Graycar, T. et al. Engineered Bacillus lentus subtilisins having altered flexibility. Journal of Molecular Biology **292**, 97-109 (1999).
8. Stroud, R.M. & Fauman, E.B. Significance of structural changes in proteins: expected errors in refined protein structures. Protein science : a publication of the Protein Society **4**, 2392-404 (1995).
9. Halle, B. Flexibility and packing in proteins. Proceedings of the National Academy of Sciences of the United States of America **99**, 1274-9 (2002).
10. Chothia, C. & Lesk, A.M. The relation between the divergence of sequence and structure in proteins. The EMBO journal **5**, 823-6 (1986).
11. Eyal, E. et al. Protein side-chain rearrangement in regions of point mutations. Proteins **50**, 272-82 (2003).
12. Eyal, E. et al. The limit of accuracy of protein modeling: influence of crystal packing on protein structure. Journal of molecular biology **351**, 431-42 (2005).
13. Brünger, A.T. Assessment of phase accuracy by cross validation: the free R value. Methods and applications. Acta crystallographica. Section D, Biological crystallography **49**, 24-36 (1993).
14. Carugo, O. & Pongor, S. A normalized root-mean-square distance for comparing protein three-dimensional structures. Protein Science: A Publication of the Protein Society **10**, 1470-3 (2001).
15. Ng, P.C. & Henikoff, S. SIFT: Predicting amino acid changes that affect protein function. Nucleic acids research **31**, 3812-4 (2003).
16. Ng, P.C. & Henikoff, S. Predicting the effects of amino acid substitutions on protein function. Annual review of genomics and human genetics **7**, 61-80 (2006).
17. Joosten, R.P. & Vriend, G. PDB improvement starts with data deposition. Science (New York, N.Y.) **317**, 195-6 (2007).
18. Yue, P., Li, Z. & Moult, J. Loss of protein structure stability as a major causative factor in monogenic disease. Journal of molecular biology **353**, 459-73 (2005).
19. Berman, H.M. et al. The Protein Data Bank. Acta crystallographica. Section D, Biological crystallography **58**, 899-907 (2002).
20. Wang, G. & Dunbrack, R.L. PISCES: a protein sequence culling server. Bioinformatics (Oxford, England) **19**, 1589-91 (2003).
21. Wang, G. & Dunbrack, R.L. PISCES: recent improvements to a PDB sequence culling server. Nucleic acids research **33**, W94-8 (2005).

22. Hamelryck, T. & Manderick, B. PDB file parser and structure class implemented in Python. Bioinformatics (Oxford, England) **19**, 2308-10 (2003).

23. The CCP4 suite: programs for protein crystallography. Acta crystallographica. Section D, Biological crystallography **50**, 760-3 (1994).

24. Hubbard, S.J. & Thornton, J.M. NACCESS.

25. R Development Core Team R: A Language and Environment for Statistical Computing. at <http://www.R-project.org>

26. Suzek, B.E. et al. UniRef: comprehensive and non-redundant UniProt reference clusters. Bioinformatics (Oxford, England) **23**, 1282-8 (2007).

Tables

Scope	Name	Information	Likely impact
Local	Maximum residue temperature factor	The largest observed atomic temperature factor for a residue	Temperature factors correlate with flexibility and thus with expected noise in structure model building
	Alpha Carbon density	The number of alpha carbon atoms within 15 Ångstroms around the alpha carbon of a given residue	Contains flexibility information beyond that in the temperature factors; can help descriptors regard to static differences between models
	Rotatable Bonds	The number of rotatable bonds in a residue's side chain	Reflects the maximum possible motion by the side-chain and contains implicit information about the size of a residue
	Crystal contacts	Whether a residue is within 4 Angstroms of another residue in the crystal lattice	Crystal contacts, like increased packing density, will correlate with lower degree of movement
	Equal crystal contact	Describes whether equivalent residues either both in or both not in a crystal contact	Captures whether or not equivalent residues are similarly constrained and may indicate conformational changes
Global	Resolution	Reflects the upper limit of resolvability of the electron density of a protein	The lower the resolution, the less well structures can be refined, therefore uncertainty is expected
	R	Self-consistency measure for assessing goodness of fit of the model with the experimental data	High R values reflect uncertainty and potential errors.
	R-free	Cross-validation measure for assessing the goodness of fit of the model with the experimental data	High R-free values reflect uncertainty and potential errors, and represent a more reliable measure than the R value.
	Mean protein temperature factor	Mean atomic temperature factor averaged over the entire structure	High global temperature factors can reflect poor data and refinement quality, and thus is expected to correlate with variation

Table 1. Survey of the descriptors used in the analysis.

	Different Author Set				Different PDB Set				Same PDB Set			
	(11299 residues from 52 pairs)				(125998 residues from 478 pairs)				(348400 residues from 1487 pairs)			
Descriptor	Estimate	Std Error	t-value	p-value	Estimate	Std Error	t-value	p-value	Estimate	Std Error	t-value	p-value
Intercept	-1.2740490	0.0354210	-35.97	<2e-16	-1.1865044	0.0118112	-100.46	<2e-16	-7.98E-01	8.41E-03	-94.86	<2e-16
Maximum residue B-factor (protein 1)	0.0269310	0.0004871	55.29	<2e-16	0.0153234	0.0001196	128.09	<2e-16	9.93E-03	9.30E-05	106.74	<2e-16
Alpha carbon density (protein 1)	-0.0110529	0.0004604	-24.00	<2e-16	-0.0114269	0.0001613	-70.82	<2e-16	-1.75E-02	1.13E-04	-154.66	<2e-16
Rotatable bonds	0.0862014	0.0053613	16.08	<2e-16	0.0947724	0.0019427	48.78	<2e-16	1.11E-01	1.34E-03	83.17	<2e-16
Residue In crystal contact (protein1)	0.4363304	0.0330958	13.18	<2e-16	0.677752	0.0127321	53.23	<2e-16	5.41E-01	7.66E-03	70.56	<2e-16
Residue in crystal contact (protein 2)	0.4043220	0.0302013	13.39	<2e-16	0.5880568	0.0129253	45.5	<2e-16	5.08E-01	7.99E-03	63.56	<2e-16
Residues in same contact state	-0.9001768	0.0465331	-19.34	<2e-16	-1.3477122	0.0187197	-71.99	<2e-16	-1.12E+00	1.14E-02	-98.06	<2e-16
Residual standard error:	0.7271 on 11292 degrees of freedom				0.8966 on 125991 degrees of freedom				1.023 on 348393 degrees of freedom			
Multiple R-squared	0.4271				0.274				0.2137			
p-value:	< 2.2e-16				< 2.2e-16				< 2.2e-16			

Table 2: Multiple regression model of the logarithmic residue RMSD against the descriptors of the purely local model.

	Different Author Set				Different PDB Set				Same PDB Set			
	(11299 residues from 52 pairs)				(125998 residues from 478 pairs)				(348400 residues from 1487 pairs)			
Descriptor	Estimate	Std. Error	t value	p-value	Estimate	Std. Error	t value	p-value	Estimate	Std. Error	t value	p-value
Intercept	-3.9802242	0.0926814	-42.945	< 2e-16	-2.9411358	0.0219473	-134.01	<2e-16	-9.91E-01	1.53E-02	-64.61	<2e-16
Worse Resolution	0.1394611	0.0250437	5.569	2.63E-08	0.3277112	0.0090182	36.34	<2e-16	-4.86E-01	5.38E-03	-90.254	<2e-16
Better Resolution	-0.4459781	0.0323169	-13.8	< 2e-16	-0.2169333	0.01077	-20.14	<2e-16	NA	NA	NA	NA
Worse R-free	1.3937611	0.3833374	3.636	0.000278	6.9826894	0.1478308	47.23	<2e-16	1.49E+01	9.97E-02	149.516	<2e-16
Better R-free	5.5477235	0.4853462	11.43	< 2e-16	4.3492844	0.1777591	24.47	<2e-16	NA	NA	NA	NA
Worse R-work	3.9400603	0.5258276	7.493	7.23E-14	-2.4945467	0.1746826	-14.28	<2e-16	-1.27E+01	1.15E-01	-110.847	<2e-16
Better R-work	3.3162877	0.5578511	5.945	2.85E-09	-2.8144235	0.1754275	-16.04	<2e-16	NA	NA	NA	NA
Worse mean B-factor	-0.0126998	0.0009276	-13.691	< 2e-16	0.0043707	0.0003275	13.35	<2e-16	9.91E-04	2.14E-04	4.633	3.61E+00
Better mean B-factor	0.0236656	0.0012781	18.517	< 2e-16	-0.015567	0.0003785	-41.13	<2e-16	NA	NA	NA	NA
Maximum residue B-factor (protein 1)	0.0244896	0.000559	43.808	< 2e-16	0.0149303	0.0001783	83.74	<2e-16	1.20E-02	1.46E-04	82.662	<2e-16
Alpha carbon density (protein 1)	-0.0098615	0.0004468	-22.073	< 2e-16	-0.0115021	0.0001571	-73.23	<2e-16	-1.63E-02	1.15E-04	-142.32	<2e-16
Rotatable bonds	0.0897666	0.0050156	17.898	< 2e-16	0.0927216	0.0018453	50.25	<2e-16	1.04E-01	1.32E-03	78.772	<2e-16
Residue In crystal contact (protein1)	0.313527	0.0306096	10.243	< 2e-16	0.5943376	0.0119025	49.93	<2e-16	5.12E-01	7.41E-03	69.056	<2e-16
Residue in crystal contact (protein 2)	0.2991043	0.0278914	10.724	< 2e-16	0.5350133	0.0120718	44.32	<2e-16	4.76E-01	7.74E-03	61.518	<2e-16
Residues in same contact state	-0.6835369	0.0431405	-15.844	< 2e-16	-1.1782788	0.0175185	-67.26	<2e-16	-1.06E+00	1.11E-02	-95.175	<2e-16
Residual standard error:	0.6684 on 11284 degrees of freedom				0.8361 on 125983 degrees of freedom				0.9887 on 348389 degrees of freedom			
Multiple R-squared	0.5162				0.3688				0.2652			
p-value:	< 2.2e-16				< 2.2e-16				< 2.2e-16			

Table 3: Multiple regression model of the logarithmic residue RMSD against the descriptors of the comprehensive model. For the global predictors, “worse” refers to the variable which is more limiting to the accuracy of the structure; for instance the higher of the two mean temperature factors (B-factors) will be classed as “worse”.

	Different Author Set	Different PDB Set	Same PDB Set	Single Point Mutant Set
Different Author Set	-	0.01526381	0.05424132	0.0722033
Different PDB Set	-	-	0.1178733	0.07107328
Same PDB Set	-	-	-	0.6966513
Single Point Mutant Set	-	-	-	-

Table 4: The significance of the differences in the distribution of the pair-wise R-squared values are tested by means of an unpaired two-tail T-test. The p-values of the all-by-all comparison are listed.

	Different Author Set				Different PDB Set				Same PDB Set			
Descriptors	Estimate	Std. Error	t value	Pr(>|t|)	Estimate	Std. Error	t value	Pr(>|t|)	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-4.84642	1.14686	-4.226	0.000122	-3.000321	0.322754	-9.296	< 2e-16	-1.17424	0.190403	-6.167	8.95E-10
Worse Resolution	0.33059	0.36264	0.912	0.367053	0.140639	0.146556	0.96	0.337738	**-0.69678**	**0.076548**	**-9.103**	**< 2e-16**
Better Resolution	-0.6543	0.41667	-1.57	0.123678	-0.353574	0.176695	-2.001	0.045964	NA	NA	NA	NA
Worse R-free	-1.72909	5.21765	-0.331	0.741958	**8.775154**	**2.473828**	**3.547**	**0.000429**	11.086562	1.38643	7.996	2.56E-15
Better R-free	4.66185	6.23723	0.747	0.458876	3.826978	2.944418	1.3	0.194329	NA	NA	NA	NA
Worse R-work	11.34963	7.55948	1.501	0.140566	-0.815344	2.951671	-0.276	0.782492	-6.926633	1.596022	-4.34	1.52E-05
Better R-work	1.53976	6.97455	0.221	0.826317	-4.307923	3.125078	-1.379	0.168706	NA	NA	NA	NA
Worse mean B-factor	-0.01058	0.0118	-0.896	0.374979	0.008566	0.005059	1.693	0.091061	0.015771	0.002166	7.282	5.34E-13
Better mean B-factor	**0.066**	**0.01762**	**3.746**	**0.000531**	0.003003	0.006331	0.474	0.63548	NA	NA	NA	NA
Residual standard error:	0.6536 on 43 degrees of freedom				0.833 on 469 degrees of freedom				0.8913 on 1482 degrees of freedom			
Multiple R-squared:	0.5361				0.1963				0.08692			
p-value:	2.519e-05				< 2.2e-16				< 2.2e-16			

Table 5: Multiple linear regression model of the logarithmic global alpha carbon variation against the global descriptors. The most informative descriptors for each set of structure pairs are in bold. As for table 3, “worse” refers to the variable which is more limiting to the accuracy of the structure; for instance the higher of the two mean temperature factors (B-factors) will be classed as “worse”.

Sample		Mean value (conservative mutants)	Mean value (non-conservative mutants)	95% confidence interval of the difference	t-value	Degrees of freedom	p-value
All	Log RMSD	-1.1115	-1.1009	-0.0511 .. 0.0297	-0.5165	7205.392	0.6056
	Z-score	-0.1481	0.1011	-0.3088 .. -0.1894	-8.1818	7247.173	3.275e-16
Top 1 residue	Log RMSD	-0.2322	-0.2302	-0.1144 .. 0.1104	-0.0354	983.939	0.9718
	Z-score	0.9563	1.2673	-0.4785 .. -0.1434	-3.6416	984.166	0.0002850
High correlation	Log RMSD	-0.4342	-1.2169	0.4473 .. 1.1181	4.6304	98.849	1.112e-05
	Z-score	-1.2143	-0.1931	-1.4434 .. -0.5990	-4.809	85.686	6.426e-06

Table 6: The difference in the distribution of logarithmic residue RMSD and Z-score values for residues in the vicinity of conservative and non-conservative mutation are listed, along with the corresponding significance values. Three samples are used: all residues pooled, the top residues per protein and all residues from protein sample, whose variation can be predicted well (R-squared value > 0.7) with the local model.

Figures

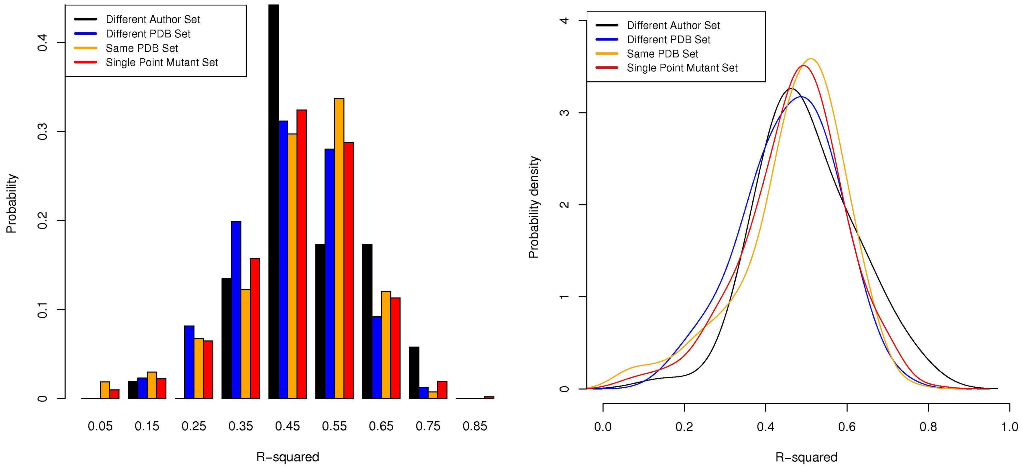


Figure 1: The distribution of multiple R-squared values as derived on a per protein pair bases. For each pair in each set, linear regression of the log RMSD versus the local descriptors was performed and the derived multiple R-squared values were pooled on a set-wise basis and plotted.

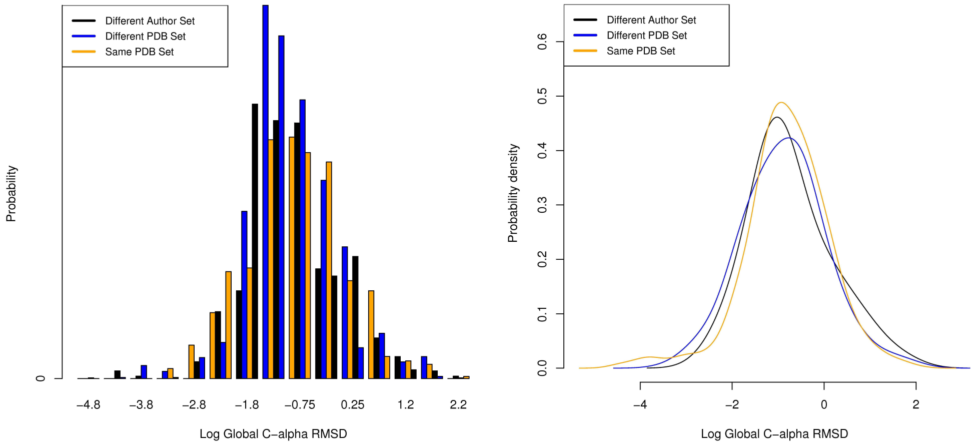


Figure 2: The logarithmic global Cα RMSD for the 3 alternative structure sets. The distribution of the histogram (right) is smoothed as outlined in M&M and the density is plotted (left).

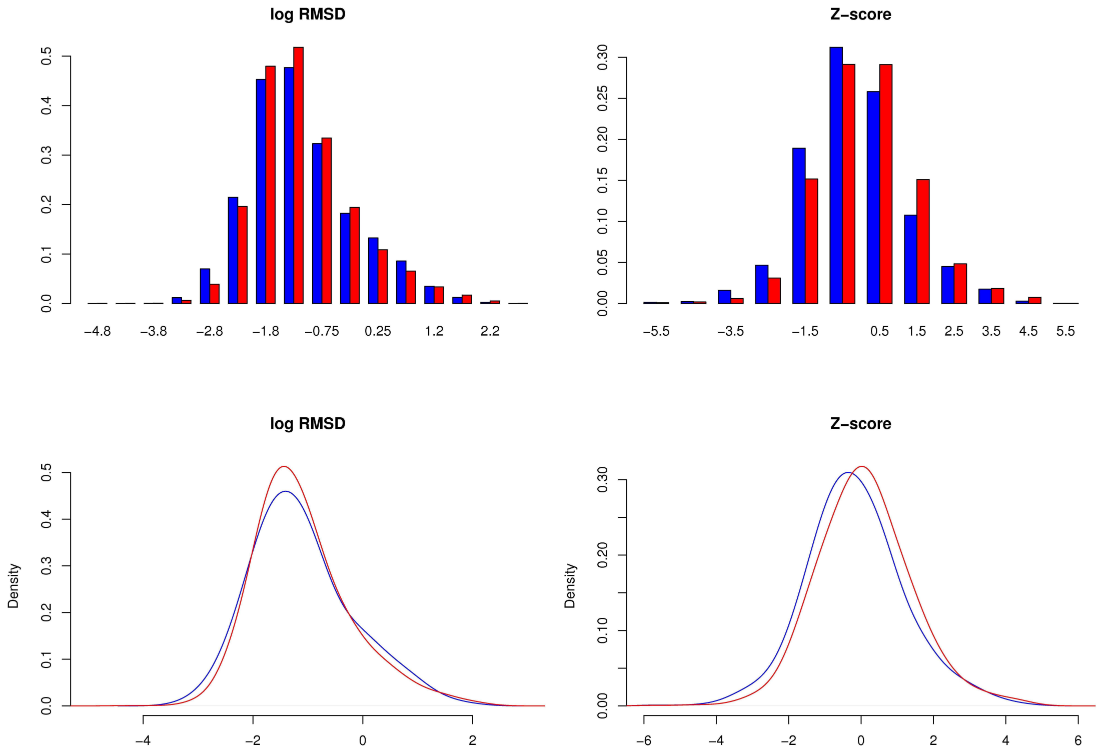


Figure 3: The difference in the distributions of log RMSD values (left hand side) and Z-scores (right hand side) for residues in the vicinity of conservative (blue) and non-conservative (red) mutation sites (see main text for definition). The histograms (upper) and density distributions (lower) show that there is a significant difference in the distributions of Z-scores, but not for raw log RMSD values. For clarity, Z-scores below -6 were set to this threshold value; this procedure was not used to calculate the T-test results in table 6.

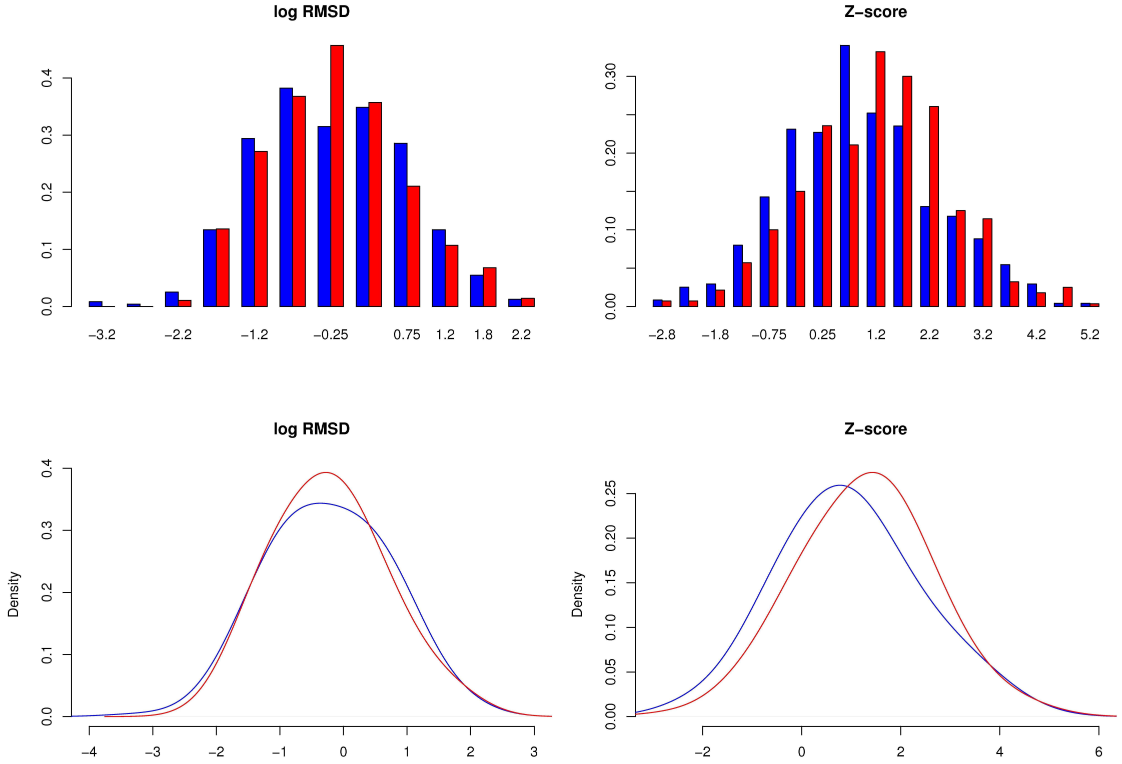


Figure 4: Histograms (upper) and density distributions (lower) of log RMSD values (left hand side) and the Z-score (right hand side) for those residues neighbouring the mutation site, which had the highest displacement by Z-score. Blue indicates that the mutations are conservative; red is used for non-conservative values.

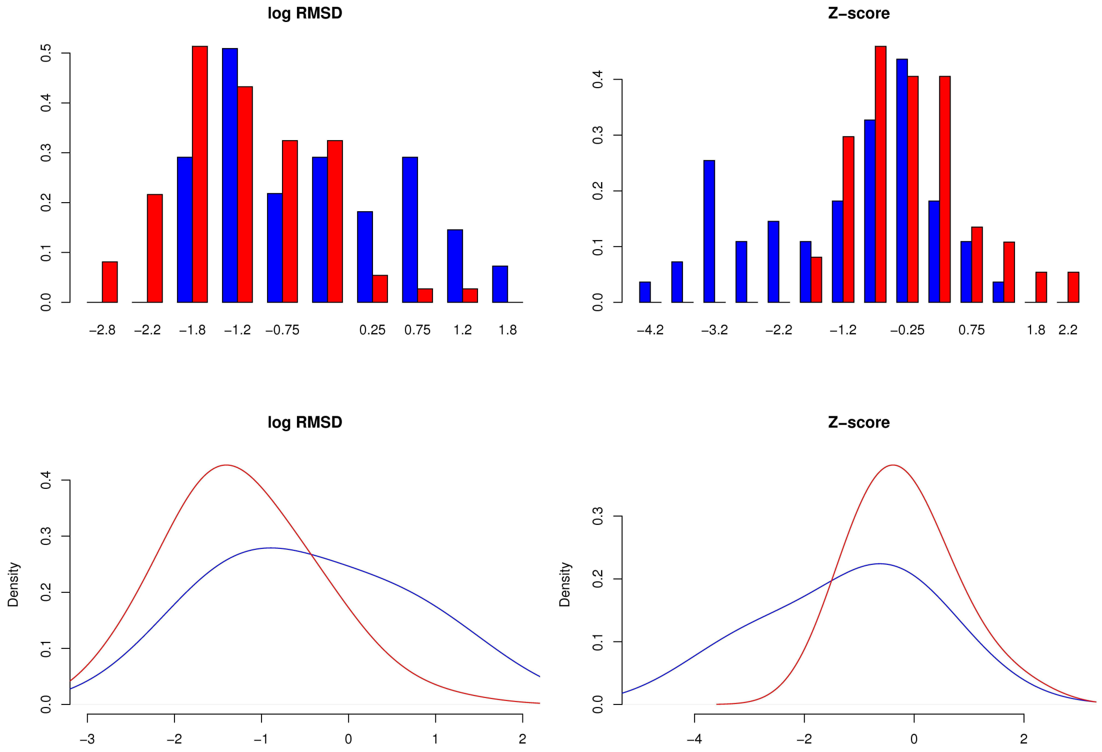


Figure 5: Histograms (upper) and density distributions (lower) of log RMSD values (left hand side) and the Z-scores (right hand side) for all residues in the vicinity of the mutation site (see main text for definition) in structures for which good correlation were achieved between residue level variability and local descriptors. Again, blue is indicative of conservative mutations, red of non-conservative ones.

2.2 Assessment of template-based predictions in the CASP7 experiment

In the following, two published manuscripts are included:

- **“Automated server predictions in CASP7”**
- **“Assessment of CASP7 predictions for template-based modeling targets”**

My contributions to the publication “Assessment of CASP7 predictions for template-based modeling targets” were as follows:

- Implementing the quality filtering steps (assigning “bumps” and “clashes”)
- Calculating all GDT and AL0 scores for ranking the models
- Performing the ranking of all models in the template-based category
- Performing the “head-to-head” t-tests of the top ranking groups

WILEY InterScience
DISCOVER SOMETHING GREAT

PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

New Folds: Server

Automated server predictions in CASP7

James N. D. Battey,[1,2] Jürgen Kopp,[1,2] Lorenza Bordoli,[1,2] Randy J. Read,[3]
Neil D. Clarke,[4] and Torsten Schwede[1,2*]

[1] Biozentrum, University of Basel, Basel, Switzerland
[2] Swiss Institute of Bioinformatics, Basel, Switzerland
[3] Department of Haematology, Cambridge Institute for Medical Research, University of Cambridge, Cambridge, United Kingdom
[4] Genome Institute of Singapore, Singapore

ABSTRACT

With each round of CASP (Critical Assessment of Techniques for Protein Structure Prediction), automated prediction servers have played an increasingly important role. Today, most protein structure prediction approaches in some way depend on automated methods for fold recognition or model building. The accuracy of server predictions has significantly increased over the last years, and, in CASP7, we observed a continuation of this trend. In the template-based modeling category, the best prediction server was ranked third overall, i.e. it outperformed all but two of the human participating groups. This server also ranked among the very best predictors in the free modeling category as well, being clearly beaten by only one human group. In the high accuracy (HA) subset of TBM, two of the top five groups were servers. This article summarizes the contribution of automated structure prediction servers in the CASP7 experiment, with emphasis on 3D structure prediction, as well as information on their prediction scope and public availability.

Proteins 2007; 69(Suppl 8):68–82.

Key words: CASP7; automated modeling server; model quality assessment.

INTRODUCTION

Protein structure prediction has become increasingly dependent on automated approaches in response to the mounting volume of data generated by large scale genome sequencing and structural genomics efforts.[1,2] Today, numerous fully automated modeling servers (for review, see recent NAR web server issue)[3] and model databases[4–6] are offering modeling services to the biomedical research community. Ultimately, the development of automated prediction methods seeks to encode expert knowledge into software. This automation allows these methods to be applied to large datasets such as whole proteomes of different organisms, as well as providing input for other prediction efforts. Since automated algorithmic approaches are devoid of human bias, their accuracy and consistency can be assessed objectively, which is an important prerequisite for their application as services for the research community.

The CASP experiment endeavors to provide a rigorous and blind assessment of state of the art methods in structure prediction.[7] Although many predictor groups participating in CASP use computational modeling procedures, these methods require human intervention at many points in the process such as performing plausibility checks, refining certain modeling steps, and selecting final models. Although this manual intervention has given human predictors a decisive advantage over fully automated methods in the past, prediction servers have played an increasingly important role. While from CASP3 to CASP5, server predictions were assessed separately as part of the CAFASP experiments,[8,9] they were part of the regular assessment as of CASP6.[10] During CASP7, prediction targets were sent to the servers automatically by the Prediction Center at UC Davis. A time limit of 48 h was imposed, in effect simulating real life modeling situations, in

The authors state no conflict of interest.

*Correspondence to: Torsten Schwede, Biozentrum University of Basel, Swiss Institute of Bioinformatics, Klingelbergstrasse 50-70, 4056 Basel, Switzerland. E-mail: torsten.schwede@unibas.ch

Received 19 May 2007; Revised 18 July 2007; Accepted 1 August 2007
Published online 25 September 2007 in Wiley InterScience (www.interscience.wiley.com).
DOI: 10.1002/prot.21761

which time is often at a premium and hence long waiting times for results are undesirable. The server predictions were then made publicly available on the CASP7 web site to allow predictor groups not registered as servers to use these data as input for their own predictions.

One aim of the CASP experiment is to measure the progress in the field. However, it has proven difficult to devise a suitable way to estimate the individual difficulty of a prediction target, which would allow comparing prediction success based on the different target data sets of two CASP experiments.[11] During the CASP7 meeting in Asilomar, it was suggested that the predictions based on servers with "frozen" algorithms using updated databases could serve as a baseline for measuring progress when comparing different CASP experiments. This approach would thereby allow the separation of improvements due to growth of underlying sequence and structure databases from those due to algorithmic developments.

This article provides an overview of all server methods participating in CASP7 in the various categories, including information on their prediction scope, public availability, URLs, and author contact information. Additionally, we summarize the results of servers in the assessment of the tertiary structure prediction categories and highlight some examples of successful predictions by servers.

SERVERS PARTICIPATING IN CASP7

In CASP7, 93 of 305 predictor groups participated as servers, with 68 servers in the tertiary structure prediction category, 8 servers in disorder prediction, 14 servers in domain boundary prediction, 8 in residue–residue contact prediction, and 6 in function prediction. Tables I and II summarize the results of a survey among the server groups regarding the following questions:

- What is the scope of the prediction server? Which input data are required?
- Is the server publicly available? Is the software available for local installation?
- Will the algorithm and/or databases be updated during the next 2 years?
- Contact details and URL for submission (if applicable).

The accuracy of server predictions in CASP7 has been assessed together with predictions submitted by manual predictor groups in each of the individual categories as described elsewhere: free modeling (FM),[66] template-based modeling (TBM),[67] high accuracy models (HA),[68] disorder,[69] domain boundary,[70] contact prediction,[71] and function prediction.[72]

Numerical assessment of the tertiary structure prediction categories in CASP7 is based on several criteria. GDT (global distance test) identifies sets of residues in the predictions deviating from the target by not more than a specified C_α distance cutoff for different sequence dependent superpositions (GDT-TS: 1, 2, 4, and 8 Å; GDT-HA: 0.5, 1, 2, and 4 Å). AL0 is defined as the percentage of correctly aligned residues in a sequence independent superposition of the model and experimental structure of the target. HBscore was introduced as an additional measure in this round of CASP in the TBM assessment. It is defined as the number of correctly predicted hydrogen bonds relative to the total number in the target structure. For this calculation we excluded side chains of residues with more than 50% relative surface exposure in the target structure, and residues with incorrect topology or involved in steric clashes in the models. For a detailed discussion of the criteria please refer to the assessment reports of the individual categories.[66–68] In the following, we will summarize the results of servers in the assessment of the tertiary structure prediction categories, and highlight some examples of successful predictions.

SERVER PREDICTIONS IN THE TBM CATEGORY

Accuracy of server predictions in the template-based modeling category

The accuracy of server predictions has continuously increased over the last years, and in CASP7 we observed a continuation of this trend with servers performing very well. In the template-based modeling (TBM) category, 68 of 187 groups were registered as prediction servers. As described in the CASP7 TBM assessment, the top 25 groups selected based on combined *z*-scores of GDT-HA and AL0 were compared by direct head-to-head comparison of statistically significant differences of GDT-HA, AL0, and HBscore on common targets.[67] Among these 25 groups, 6 were servers with the best group (25 Zhang-server) ranked third overall.

Here, we aimed at a direct comparison of only servers in this category. We have therefore recalculated the numerical assessment, taking only into account the predictions submitted by servers. The results are presented in Figure 1 and Table III. The best performing group Zhang-server (group 25) is followed by servers developed by Soeding *et al.* (213 HHpred2; 214 BayesHH; 418 HHpred3), Elofsson and coworkers (47 Pmodeller6), Baker and coworkers (4 Robetta), and Skolnick and coworkers (307 MetaTasser).

Examples of successful server predictions

Several examples of outstanding predictions submitted by servers were observed in CASP7, e.g. for target T0321 (PDB:2h1q), which is a structural genomics target from *Desulfitobacterium hafniense* of unknown function. The protein of 250 amino acid residues forms a two-domain mixed αβ-structure. The C-terminal domain is character-

Table I
Servers Participating in CASP7

Server	CASP7 ID	Which type of prediction does this server offer?	Is the server publicly available?	Which input data are required?	Which output data are generated?	Average server response time	CASP8	Method updates	Data updates	Local installation	Reference
3D-JIGSAW	302	TBM	Y	S, A	FA, TTA	∼30 min	Y	Y	Y	N	12
3D-JIGSAW_POPULUS	247	TBM, FM	N	–^{c}	–^{c}	–^{c}	–^{c}	Y	Y	Y	13
3D-JIGSAW_RECOM	420	TBM	Y	C	FA, energy curve	∼1 h	Y	N	N	N	14
3Dpro	137	FM	Y	S	FA	∼10 min	Y	Y	–^{c}	N	15
ABIpro	139	FM	Y	S	FA	∼3 h	Y	N	Y	N	15
BAKER-ROSETTADOM	497	DBP	N	–^{c}	–^{c}	–^{c}	–^{c}	–^{c}	–^{c}	N	16
BayesHH	214	–^{c}	–^{c}	–^{c}	–^{c}	–^{c}	–^{c}	–^{c}	–^{c}	–^{c}	–^{c}
beautshot	275	TBM	N	S	FA	∼2 min	Y	Y	Y	N	–^{c}
Beautshotbase	347	TBM	N	–^{c}	–^{c}	–^{c}	–^{c}	–^{c}	–^{c}	N	–^{c}
BETApro	141	CP	Y	S	–^{c}	Several minutes	Y	–^{c}	–^{c}	Y	17
bicmkusk-serv	202	–^{c}	–^{c}	–^{c}	–^{c}	–^{c}	–^{c}	–^{c}	–^{c}	–^{c}	–^{c}
Bilab-ENABLE	179	TBM	N	S	FA	∼3 h	Y	Y	Y	N	18
BIME@NTU_serv	272	DP	Y	S	Propensity for disorder/order	∼5 min	Y	Y	Y	N	19
Casplta-FOX	186	TBM	Y	S	FA	∼14 h	Y	Y	Y	N	–^{c}
Casplta-GOret	573	FP	Y	S	Prediction sent by email in Casp7 FN format	∼1 h	Y	Y	Y	N	–^{c}
Chop	595	DBP	–^{c}	–^{c}	Neural network prediction of domain linking regions	–^{c}	–^{c}	–^{c}	–^{c}	–^{c}	–^{c}
Chop_homo	649	DBP	–^{c}	–^{c}	Homology based method for domain prediction	–^{c}	–^{c}	–^{c}	–^{c}	–^{c}	–^{c}
CIRCLE	298	TBM	N	S	FA	∼24 h	Y	Y	Y	N	–^{c}
CPHmodels	494	TBM	Y	S	FA	∼1 min	Y	Y	Y	Y	20
DISOPRED	470	DP	Y	S	Regions predicted to be disordered	–^{c}	Y	N	–^{c}	Y	21
DISpro	140	DP	Y	–^{c}	–^{c}	–^{c}	–^{c}	–^{c}	–^{c}	Y	22
Distill	168	TBM, FM, DP, DBP	Y	S	CA, contact maps, SS, solvent accessibility, etc.	∼10 min	Y	Y	Y	Y	23
DomFOLD	240	DBP	Y	S	Domain prediction in CASP DP format	∼30 min	Y	Y	Y	N	–^{c}
DomSSEA	312	DBP	Y	S	Domain boundary prediction	–^{c}	Y	–^{c}	–^{c}	N	24
DPS	310	DBP	Y	S	Domain boundary prediction	–^{c}	Y	–^{c}	–^{c}	N	24
DRIPPRED	153	DP	Y	S	Per-residue disorder score	∼1 h	Y	N	Y	N	–^{c}
FAMS	351	TBM, FM	Y	S	FA	∼7 days	Y	Y	Y	N	25
FAMSD	349	TBM	Y	S	FA	∼2 days	Y	N	Y	N	25
FOLDpro	136	TBM	Y	S	FA, TTA	Several hours	Y	Y	Y	N	26
forecast-s	333	–^{c}	–^{c}	–^{c}	–^{c}	–^{c}	–^{c}	–^{c}	–^{c}	–^{c}	–^{c}
FORTE1	257	TBM	Y	S	FA, TTA	–^{c}	Y	–^{c}	Y	N	27
FORTE2	316	TBM	N	–^{c}	–^{c}	–^{c}		–^{c}	Y	N	28
FPSOLVER-SERVER	511	FM	–^{c}	S	FA	∼2 days	N	Y	N	N	–^{c}
Frankenstein	368	TBM	Y	S	FA	∼3 days	Y	Y	Y	N	29
FUGMOD	319	TBM	N	S, A	FA	–^{c}	N	–^{c}	–^{c}	N	30
FUGUE	242	TBM	Y	S, A	TTA, BB (without loops)	∼10 min	Y	–^{c}	Y	Y	30
FUNCTION	318	TBM	N	–^{c}	FA	∼24 h	N	Y	Y	N	–^{c}

(*Continued*)

Table I
(Continued)

Server	CASP7 ID	Which type of prediction does this server offer?	Is the server publicly available?	Which input data are required?	Which output data are generated?	Average server response time	CASP8	Method updates	Data updates	Local installation	Reference
GajdaPairings	618	CP	Y	S	Robust contact prediction	~3 days	Y	Y	Y	N	—^{c}
GeneSilicoMetaServer	609	TBM, DBP, DP, MQE, SSP	Y	S, A	FA, TTA	~15 min	Y	Y	Y	N	31
GeneSilicoUnimod	461	—^{c}	—^{c}	—^{c}	—^{c}	—^{c}	—^{c}	—^{c}	—^{c}	—^{c}	—^{c}
GPCPRED	154	CP	Y	S	CASP format contact prediction, 2D contact map image in HTML version	~60 min	Y	N	Y	N	32
Gtg	44	TBM	Y	S, DBI	FA	Seconds	Y	N	Y	Y	—^{c}
HHpred1	212	TBM, DBP, FP	Y	S	Confidence value for homology, TTA, FA; Domain prediction; GO function prediction	~10 min	Y	Y	Y	Y	33
HHpred2	213	TBM	Y	S	Confidence value for homology, TTA, FA; domain prediction; GO function prediction	~12 min	Y	Y	Y	Y	33
HHpred3	418	TBM, DBP, FP	Y	S	Confidence value for homology, TTA, FA; domain prediction; GO function prediction	~17 min	N	Y	Y	N	33
Huber-Torda-Server	102	TBM	Y	S	Ranked list of templates, models complete only up to beta carbons	~75 min	Y	Y	Y	Y	34
karypis.srv	22	TBM	Y	S	FA	~10 min	Y	Y	Y	N	—^{c}
karypis.srv.2	268	TBM, MQE	Y	S	FA	~2 days	N	Y	N	Y	35,36
karypis.srv.4	193	FM	Y	S	FA	~30 min	N	N	N	N	—
keasar-server	277	TBM	Y	S	FA	~1 h	Y	Y	N	Y	37
LOOPP	83	TBM	Y	S	FA, TTA	~5 h	Y	Y	Y	N	38–40
Ma-OPUS-DOM	229	DBP	Y	S	Domain information in CASP format	~1.5 days	Y	Y	Y	N	—^{c}
Ma-OPUS-server	92	TBM, FM	Y	S	FA	~5 h	Y	Y	Y	N	—^{c}
Ma-OPUS-server2	728	TBM, FM	Y	S	FA	~5 h	Y	Y	Y	N	—^{c}
MBI-NTU-serv	538	—^{c}	—^{c}	—^{c}	—^{c}	—^{c}	—^{c}	—^{c}	—^{c}	—^{c}	—^{c}
Meta-DP	269	DBP	Y	S	Domains and domain boundaries	~5 min	Y	N	Y	N	41
MetaTasser	307	TBM, FM	Y	S	FA	—^{c}	Y	Y	Y	Y	42–44
mGen-3D	261	TBM	N	—^{c}	—^{c}	—^{c}	—^{c}	—^{c}	—^{c}	—^{c}	—^{c}
MIG_FROST	586	—^{c}	—^{c}	—^{c}	—^{c}	—^{c}	—^{c}	—^{c}	—^{c}	—^{c}	—^{c}
MIG_FROST_FLEX	588	—^{c}	—^{c}	—^{c}	—^{c}	—^{c}	—^{c}	—^{c}	—^{c}	—^{c}	—^{c}
nFOLD	239	TBM	Y	S	BB	~30 min	Y	N	Y	N	45
NN_PUT_lab	654	TBM, DBP	Y	S	FA, metaserver	—^{c}	Y	Y	Y	N	—^{c}
panther2	69	—^{c}	—^{c}	—^{c}	—^{c}	—^{c}	—^{c}	—^{c}	—^{c}	—^{c}	—^{c}
panther3	304	—^{c}	—^{c}	—^{c}	—^{c}	—^{c}	—^{c}	—^{c}	—^{c}	—^{c}	—^{c}

(Continued)

Table I
(Continued)

Server	CASP7 ID	Which type of prediction does this server offer?	Is the server publicly available?	Which input data are required?	Which output data are generated?	Average server response time	CASP8	Method updates	Data updates	Local installation	Reference
Pcons6	46	TBM, FM, MQE	Y	S	FA, ALN, QS	5–30 min	Y	Y	Y	Y	46
PFP_HAWKINS	753	FP	Y	S	Predicted Gene Ontology annotations (MF, BP, CC)	~30 min	Y	Y	Y	Y	47
Phyre-1	468	TBM	Y	S	FA	30 min–1 h	Y	–[c]	N	Y	s
Phyre-2	469	TBM, FM	N	S	FA	~10 h	N	Y	Y	N	–[c]
Pmodeller6	47	TBM, FM, MQE	Y	S	FA, ALN, QS	5–30 min	Y	Y	Y	Y	46
POMYSL	464	–[c]	–[c]	–[c]	–[c]	–[c]	–[c]	–[c]	–[c]	–[c]	–[c]
Possum	230	CP	Y	S	Contact pairs in Casp format	~5 min	Y	Y	Y	N	48
PROFcon-Rost	296	CP	Y	S	List of amino acid pairs ranked according to PROFcon contact score (low values = contact unlikely)	1 h	Y	N	Y	N	49
PROTINFO	28	TBM, FM	Y	–[c]	–[c]	–[c]	Y	Y	Y	Y	50
PROTINFO-AB	29	TBM, FM[b]	Y	S, A	Depends on input	3–24 h	Y	Y	Y	Y	50
Raghava-GPS-mango	598	FP	Y	S	Text	~50 s	Y	Y	Y	N	51
RAPTOR	248	TBM	Y	S	FA	~5 h	Y	Y	Y	N	52
RAPTOR-ACE	267	TBM	N	–[c]	–[c]	–[c]	Y	Y	Y	Y	53
RAPTORESS	435	TBM	N	S	FA	–[c]	N	Y	Y	N	–[c]
ROBETTA	4	TBM, FM, DBP, FL, AS	Y	S	FA, TTA, MAMMOTH hits to known structures for de novo models	~1 month	Y	N	Y	N	54
ROBETTA-GINZU	581	DBP	Y	S	Domain boundary prediction	2 h	Y	N	Y	N	16,54
ROKKY	35	–[c]	–[c]	–[c]	–[c]	–[c]	–[c]	–[c]	–[c]	–[c]	–[c]
Rost_PROFbval	594	Bval	Y	S, MSA	Per residue normalized B-value; 2 state prediction - flexible/rigid	–[c]	Y	N	–[c]	N	55,56
Rost-ECGO	751	FP	N	S	Predicted EC class	5 min	Y	Y	N	N	–[c]
SAM_T06_server	389	TBM, FM, CP, LSP	Y	S	FA, TTA, residue-residue contact predictions, local structure predictions, MSA, HMMs, . . .	~ 12 h	Y	N	Y	N	57
SAM-T02[a]	381	TBM, SSP, MSA	Y	S	TTA, secondary structure predictions, burial predictions	~ 4 h	Y	N	Y	N	58
SAM-T99[a]	380	TBM, SSP, MSA	Y	S	TTA, MSA, SS	~ 12 h	Y	N	Y	N	59
Shub	274	TBM	N	–[c]	–[c]	–[c]	N	Y	N	N	–[c]
SP3	414	TBM	Y	S	FA, TTA	Several hours	Y	N	Y	Y	60
SP4	415	TBM	Y	S	FA, TTA	Several hours	Y	N	Y	Y	61
SPARKS2	413	TBM	Y	S	FA, TTA	Several hours	Y	N	Y	Y	60
SVMcon	138	CP	N	–[c]	–[c]	–[c]	–[c]	–[c]	–[c]	–[c]	62
UNI-EID_bnmx	383	TBM	N	–	–	–	–	–	–	N	63
UNI-EID_expm	245	TBM	N	–	–	–	–	–	–	N	64

(*Continued*)

Table I
(Continued)

Server	CASP7 ID	Which type of prediction does this server offer?	Is the server publicly available?	Which input data are required?	Which output data are generated?	Average server response time	CASP8	Method updates	Data updates	Local installation	Reference
UNI-EID_sfst	243	TBM	N	–	–	–	–	–	–	N	63
Zhang-Server	25	TBM, FM	Y	S	FA	~10 h	Y	Y	Y	N	65

The presented information about the participating servers was collected in a survey among the registered groups after the experiment. The table summarizes group name and number, type of predictions performed, public accessibility of the server, required input data and output format, and average response time. The rightmost columns indicate if the server is expected to be available at least until CASP8, if the algorithm and underlying databases will be modified in the meantime, and if the software is available for local instillation. For specific comments provided by the server authors see footnotes.

TBM, template-based modeling; FM, free modeling; DBP, domain boundary prediction; DP, disorder prediction; CP, contact prediction; FP, function prediction; SSP, secondary structure prediction; LSP, local structure prediction; MSA, multiple sequence alignment; MQE, model quality estimation; Bval, normalized B values; FL, fragment libraries; AS, computational Alanine scanning; S, protein sequence; ALN, sequence alignment; C, coordinates; DBI, database identifier; FA, full atom model; BB, backbone model; CA, alpha carbon trace; TTA, target template alignment; SS, secondary structure prediction; MSA, multiple sequence alignment; QS, quality score; s, manuscript submitted for publication.

[a] This server is obsolete and is being kept alive only for historical comparisons.

[b] Optional experimental data.

[c] No information provided by authors.

ized by an extended central β-sheet flanked by four α-helices and has been classified as TBM/FM prediction target since a significant part of the structure could not be modeled based on the available template structure. However, for the N-terminal domain, one server (415 SP4) recognized the structural similarity to the N-terminal domain of Enolases (CATH code 3.30.390.10).[73] The C-terminal domains of enolases are TIM barrels (CATH code 3.20.20.120) and do not resemble the second domain of target T0321. The submitted model by the SP4 server for domain 1 based on mandelate racemase from *Pseudomonas putida* (PDB:2mnr) as template (Fig. 2) achieved a GDT-HA of 35.2 (AL0 of 49.0), which is outstanding when compared with a GDT-HA of 24.2 (AL0 of 0.0) of the second best prediction.

Target T0356 is also a structural genomics target, the 3-octaprenyl-4-hydroxybenzoate decarboxylase (UbiD) from *Escherichia coli* (PDB:1idb). For the assessment, T0356 has been divided into three assessment units: domains 1 and 3 were assessed in the FM category; the second domain, which resembled an FMN-binding protein domain (CATH code 2.30.110.10), was assessed as a TBM target. The best available template, the structure of an archeal FMN-binding protein from *Methanobacterium thermoautotrophicum* (PDB:1eje), was used for several of the best-submitted predictions. The structural similarity was difficult to detect, and only predictions by eight groups were significantly better than the remainder (Fig. 3)—among which seven were registered as servers (212 HHpred1; 213 HHpred2; 214 BayesHH; 418 HHpred3; 92 Ma-OPUS-server; 245 UNI-EID_expm; 383 UNI-EID_bnmx). Interestingly, only one metapredictor method (675 Fams-ace), and none of the manual predictor groups made use of these server predictions.

Limitations in template detection

TBM exploits the evolutionary relationship between a target and a template protein to infer structural similarity. In cases of high sequence identity between the target and the template, simple algorithms for sequence alignment are sufficient for identifying and aligning the best template to the target. If the similarity is low, the detection and alignment of templates require more sophisticated methods. A good template may exist for a target, yet not be detectable by simple sequence-based methods. Fold-recognition methods attempt to address the problem of detecting such remote homologs. As illustrated in the previous two examples, this problem is still far from being generally solved, and considerable performance differences can be attributed to the ability of servers to build their models on the best available templates. Here, we sought to address two issues. The first is whether a server was able to detect the best possible structural template and the second is how well it would have fared in

Table II
Contact Details for Publicly Available CASP7 Servers

Server	URL	Contact E-mail
3D-JIGSAW	http://www.bmm.icnet.uk/∼3djigsaw/	paul.bates@cancer.org.uk
3D-JIGSAW_RECOM	http://www.bmm.icnet.uk/servers/3djigsaw/recomb/index.html	
3Dpro	http://www.ics.uci.edu/∼baldig/scratch/	pfbaldi@ics.uci.edu
ABIpro	http://www.ics.uci.edu/∼baldig/scratch/	arandall@ics.uci.edu
BETApro	http://www.igb.uci.edu/?page=tools&subPage=psss	pfbaldi@ics.uci.edu
BIME@NTU_serv	http://biominer.bime.ntu.edu.tw/casp7/	cychen@mars.csie.ntu.edu.tw
CaspIta-FOX	http://protein.cribi.unipd.it/fox/	stefano.toppo@unipd.it
CaspIta-GOret	http://protein.cribi.unipd.it/go_retriever/	
CPHmodels	http://www.cbs.dtu.dk/services/CPHmodels/	lund@cbs.dtu.dk
DisoPred	http://bioinf.cs.ucl.ac.uk/disopred/	d.jones@cs.ucl.ac.uk
DISpro	http://www.ics.uci.edu/∼baldig/scratch/	pfbaldi@ics.uci.edu
Distill	http://distill.ucd.ie/distill/	gianluca.pollastri@ucd.ie
DomFOLD	http://www.biocentre.rdg.ac.uk/bioinformatics/DomFOLD/DomFOLD_form.html	l.j.mcguffin@reading.ac.uk
DomSSEA	http://bioinf.cs.ucl.ac.uk/dompred/	k.bryson@cs.ucl.ac.uk
DPS	http://bioinf.cs.ucl.ac.uk/dompred/	
DRIPPRED	http://sbcweb.pdc.kth.se/cgi-bin/maccallr/disorder/submit.pl	r.maccallum@imperial.ac.uk
FAMS	http://www.pharm.kitasato-u.ac.jp/fams/fams.html	kanouk@pharm.kitasato-u.ac.jp
FAMSD	http://www.pharm.kitasato-u.ac.jp/fams/famsd.html	
FOLDpro	http://mine5.ics.uci.edu:1026/foldpro.html	pfbaldi@ics.uci.edu
FORTE1	http://www.cbrc.jp/forte/	k-tomii@aist.go.jp
Frankenstein	https://genesilico.pl/meta2	mgajda@genesilico.pl
FUGUE	http://tardis.nibio.go.jp/fugue/; http://www-cryst.bioc.cam.ac.uk/fugue/	kenji@nibio.go.jp
GajdaPairings	https://genesilico.pl/meta2	mgajda@genesilico.pl
GeneSilicoMetaServer	https://genesilico.pl/meta2	andrzej@genesilico.pl
GPCPRED	http://sbcweb.pdc.kth.se/cgi-bin/maccallr/gpcpred/submit.pl	r.maccallum@imperial.ac.uk
gtg	http://www.bioinfo.biocenter.helsinki.fi/gtg	liisa.holm@helsinki.fi
HHpred1	http://protevo.eb.tuebingen.mpg.de/∼toolkit/hhpred1/	johannes.soeding@tuebingen.mpg.de
HHpred2	http://protevo.eb.tuebingen.mpg.de/∼toolkit/hhpred2/	
HHpred3	http://protevo.eb.tuebingen.mpg.de/∼toolkit/hhpred3/	
Huber-Torda-Server	http://www.zbh.uni-hamburg.de/wurst/	torda@zbh.uni-hamburg.de
karypis.srv	http://www.cs.umn.edu/∼karypis/servers/c7pred	karypis@cs.umn.edu
karypis.srv.2	http://dminers.dtc.umn.edu/∼rangwala/mn-fold/fp.php	rangwala@cs.umn.edu
karypis.srv.4	http://www-users.cs.umn.edu/∼deronne/c7pred/	deronne@cs.umn.edu
keasar-server	http://www.cs.bgu.ac.il/∼meshisrv/server/	keasar@cs.bgu.ac.il
LOOPP	http://cbsuapps.tc.cornell.edu/loopp.aspx	ron@cs.cornell.edu
Ma-OPUS-server	http://sigler.bioch.bcm.tmc.edu/MaLab/CASP7-server/	jpma@bcm.tmc.edu
Ma-OPUS-server2	http://sigler.bioch.bcm.tmc.edu/MaLab/CASP7-server2/	
Ma-OPUS-DOM	http://sigler.bioch.bcm.tmc.edu/CASP7-DOM/	
Meta-DP	http://meta-dp.cse.buffalo.edu	hksaini@cse.buffalo.edu
MetaTasser	http://cssb.biology.gatech.edu/skolnick/webservice/MetaTASSER/	skolnick@gatech.edu
nFOLD	http://www.biocentre.rdg.ac.uk/bioinformatics/nFOLD/nFOLD_form.html	l.j.mcguffin@reading.ac.uk
NN_PUT_lab	http://webmobis.cs.put.poznan.pl	protserv@cs.put.poznan.pl
Pcons6	http://pcons.net	bjorn@sbc.su.se
PFP_HAWKINS	http://dragon.bio.purdue.edu/pfp	thawkins@purdue.edu
Phyre-1	http://www.sbg.bio.ic.ac.uk/∼phyre/	l.a.kelley@imperial.ac.uk
Pmodeller6	http://pcons.net	bjorn@sbc.su.se
Possum	http://foo.maths.uq.edu.au/∼nick/Protein/contact.html	n.hamilton@imb.uq.edu.au
PROFcon-Rost	http://www.predictprotein.org/submit_profcon.html	mp2215@columbia.edu
PROTINFO	http://protinfo.compbio.washington.edu/protinfo_abcmfr/	ram@compbio.washington.edu
PROTINFO-AB	http://protinfo.compbio.washington.edu	admin@protinfo.compbio.washington.edu
Raghava-GPS-mango	http://www.imtech.res.in/raghava/mango/	raghava@imtech.res.in
RAPTOR	http://ttic.uchicago.edu/∼jinbo/	j3xu@tti-c.org
ROBETTA	http://robetta.org/submit.jsp	DCChivian@lbl.gov
ROBETTA-GINZU	http://robetta.org/submit.jsp	
Rost-ECGO	http://rostlab.org/services/ecgo/	amk2002@columbia.edu
Rost_PROFbval	http://rostlab.org/services/profbval/	as2067@columbia.edu
SAM_T06_server	http://www.soe.ucsc.edu/research/compbio/SAM_T06/T06-query.html	sam-info@soe.ucsc.edu
SAM-T02[a]	http://www.soe.ucsc.edu/research/compbio/SAM_T02/T02-query.html	
SAM-T99[a]	http://www.soe.ucsc.edu/research/compbio/HMM-apps/T99-query.html	
SP3	http://sparks.informatics.iupui.edu	yqzhou@iupui.edu
SP4	http://sparks.informatics.iupui.edu	
SPARKS2	http://sparks.informatics.iupui.edu	
Zhang-Server	http://zhang.bioinformatics.ku.edu/I-TASSER	yzhang@ku.edu

The presented information about the participating servers was collected in a survey among the registered groups after the experiment.

[a]This server is obsolete and is being kept alive only for historical comparisons.

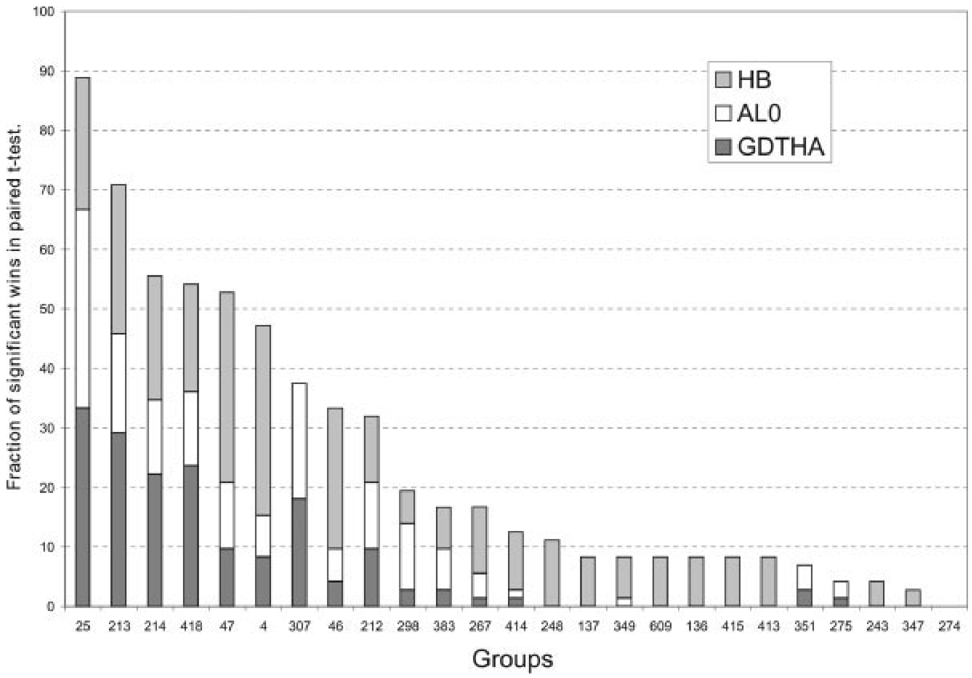


Figure 1

Head-to-head comparison for the top 25 server groups showing the fraction of statistically significant wins (Student's t*-test;* P*-value < 0.05) on common targets.*

comparison with the best template identified using a purely sequence-based method.

The best-possible template is detected using a structure based search of the target against the PDB entries available up until the target's submission deadline as described elsewhere in this issue.[74] Specifically, the highest scoring structure according to the sequence-independent superposition generated by LGA[75] was defined to be the best template. "Pseudopredictions" were built based on LGA's structural alignment using a 4 Å superposition cutoff, whereby the coordinates of aligned residues were copied from the equivalent residues in the template.

For comparison, we used PSI-BLAST[76] to identify and align a template to the target sequence. The initial PSI-BLAST profile was generated for the target sequence on the NCBI nonredundant protein sequence database[77] and subsequently used to scan the PDB for templates available during the prediction window.

Using the lowest e-value as a criterion for choosing PSI-BLAST hits frequently identifies short fragments of very high similarity, but which give rise to models with a low GDT-HA due to the low coverage of the target. In the case of multidomain proteins, the e-value calculated on the basis of the whole sequence is not applicable for the individual subunits. Therefore, we filtered the PSI-BLAST hits by maximal coverage of the individual assessment units, and subsequently chose the highest ranked PSI-BLAST hits by e-value. Pseudopredictions were built by copying the backbone coordinates for all residues aligned between target and template in the PSI-BLAST alignment. No attempt was made to further improve the alignment or to model insertions or deletions.

For 41% of targets (44 of 108), the pseudopredictions built using the PSI-BLAST templates are within 10 GDT-HA units of the best structural template available. In these cases, the simple sequence search correctly identified a suitable structural template and little improvement would have been possible. For the remaining 64 targets, however, considerable improvement of prediction quality over that of the PSI-BLAST template would have been possible, namely by a margin of more than 10 GDT-HA points in 64 cases and even 20 points in 41 cases. However, it must be pointed out that for 30 of these 41 cases, the pseudopredictions based on the optimal structural template identified by LGA have GDT-HA scores below 50. These models are either incomplete or the templates are structurally divergent from the target.

We compared the performance of both sets of pseudopredictions, "PSI-Blast template" and "LGA template," to that of the best of all submitted server models, the best overall server (25 Zhang-server) and the best metapredictor server (47 Pmodeller6). Using the PSI-BLAST based model as baseline, we subtracted its GDT-HA value from that of the other predictions for each target and plotted the results in Figure 4 (upper panel). The Zhang-server

Table III

Statistical Significance of the Results of the 25 Highest Scoring Server Groups

	25	213	418	307	214	47	212	4	298	46	383	351	137	267	349	248	414	609	275	136	415	243	413	347	274
25		1.31(0.02) 3.20(0.00)	1.78(0.00) 3.68(0.00)	2.25(0.00) 3.06(0.00)	2.03(0.00) 4.10(0.00)	2.54(0.00) 4.36(0.00)	2.53(0.00) 4.09(0.00)	2.65(0.00) 4.47(0.00)	3.13(0.00) 4.24(0.00)	2.97(0.00) 4.98(0.00)	3.23(0.00) 4.78(0.00)	3.24(0.00) 5.09(0.00)	3.59(0.00) 6.82(0.00)	3.40(0.00) 5.25(0.00)	3.52(0.00) 5.98(0.00)	4.08(0.00) 6.43(0.00)	3.47(0.00) 6.17(0.00)	3.97(0.00) 6.45(0.00)	3.39(0.00) 4.76(0.00)	3.67(0.00) 7.18(0.00)	3.82(0.00) 6.64(0.00)	4.00(0.00) 6.16(0.00)	4.08(0.00) 7.01(0.00)	3.77(0.00) 7.10(0.00)	4.43(0.00) 6.75(0.00)
213	0.15(0.86) n = 108		0.47(0.03) 0.49(0.30)	0.98(0.10) -0.05(0.96)	0.55(0.16) 0.71(0.33)	1.27(0.02) 1.19(0.23)	1.22(0.00) 0.90(0.14)	1.36(0.01) 1.25(0.17)	1.84(0.00) 1.21(0.20)	1.64(0.00) 1.70(0.06)	1.56(0.00) 1.07(0.22)	1.92(0.00) 2.06(0.03)	2.27(0.00) 3.59(0.00)	2.03(0.00) 2.00(0.06)	2.05(0.00) 2.31(0.02)	2.77(0.00) 3.23(0.01)	2.20(0.00) 3.04(0.01)	2.36(0.00) 2.90(0.01)	1.75(0.00) 0.85(0.34)	2.36(0.00) 3.98(0.00)	2.52(0.00) 3.44(0.00)	2.20(0.00) 2.09(0.02)	2.78(0.00) 3.81(0.00)	2.38(0.00) 3.72(0.00)	2.74(0.00) 2.61(0.02)
418	0.74(0.39) n = 108	0.58(0.04) n = 108		0.49(0.38) -0.45(0.65)	0.09(0.79) 0.23(0.71)	0.81(0.16) 0.74(0.48)	0.75(0.09) 0.41(0.54)	0.88(0.11) 0.76(0.41)	1.35(0.01) 0.72(0.44)	1.16(0.04) 1.21(0.18)	1.18(0.04) 0.89(0.36)	1.43(0.01) 1.57(0.10)	1.79(0.00) 3.09(0.00)	1.55(0.01) 1.51(0.17)	1.55(0.01) 1.89(0.07)	2.29(0.00) 2.75(0.03)	1.72(0.00) 2.55(0.01)	1.92(0.00) 2.51(0.01)	1.42(0.01) 0.54(0.51)	1.89(0.00) 3.50(0.00)	2.04(0.00) 2.96(0.01)	1.85(0.00) 1.81(0.05)	2.30(0.00) 3.33(0.00)	1.93(0.00) 3.20(0.00)	2.22(0.00) 2.21(0.03)
307	5.31(0.00) n = 104	5.27(0.00) n = 104	4.63(0.00) n = 104		-0.31(0.57) 0.83(0.36)	0.44(0.39) 1.47(0.06)	0.29(0.62) 1.01(0.36)	0.54(0.31) 1.43(0.13)	0.90(0.10) 1.29(0.17)	0.85(0.13) 1.78(0.07)	0.83(0.18) 1.59(0.10)	0.91(0.08) 2.01(0.01)	1.45(0.01) 3.91(0.00)	1.14(0.03) 2.11(0.01)	1.28(0.03) 2.83(0.00)	1.93(0.00) 3.31(0.00)	1.28(0.01) 3.09(0.00)	1.64(0.00) 3.33(0.00)	1.15(0.02) 1.82(0.02)	1.50(0.02) 4.17(0.00)	1.59(0.00) 3.54(0.00)	1.56(0.01) 2.82(0.00)	1.90(0.00) 3.91(0.00)	1.55(0.01) 4.05(0.00)	2.17(0.00) 3.41(0.00)
214	0.32(0.71) n = 101	0.28(0.62) n = 101	-0.28(0.59) n = 101	**-4.76(0.00)** n = 97		0.72(0.20) 0.42(0.65)	0.78(0.17) 0.45(0.62)	0.57(0.28) 0.17(0.83)	1.16(0.01) 0.37(0.59)	1.08(0.05) 0.69(0.39)	1.32(0.04) 1.00(0.33)	1.16(0.02) 1.17(0.16)	1.47(0.02) 2.38(0.01)	1.42(0.02) 1.24(0.22)	1.45(0.01) 1.61(0.10)	2.18(0.00) 2.54(0.02)	1.80(0.00) 2.32(0.01)	1.90(0.00) 2.38(0.02)	1.25(0.01) 0.20(0.75)	1.63(0.01) 2.95(0.00)	1.83(0.00) 2.65(0.01)	1.92(0.00) 1.81(0.06)	2.17(0.00) 3.04(0.00)	1.93(0.00) 3.25(0.00)	2.29(0.00) 2.39(0.03)
47	**-1.95(0.01)** n = 107	**-1.98(0.01)** n = 107	**-2.56(0.00)** n = 107	**-7.04(0.00)** n = 103	**-2.11(0.01)** n = 100		-0.07(0.90) -0.38(0.73)	0.05(0.90) -0.03(0.97)	0.56(0.27) -0.01(0.99)	0.46(0.32) 0.51(0.56)	0.41(0.48) 0.11(0.91)	0.59(0.20) 0.81(0.27)	0.92(0.12) 2.26(0.02)	0.81(0.14) 0.81(0.37)	0.84(0.08) 1.36(0.08)	1.48(0.01) 1.96(0.06)	0.91(0.07) 1.79(0.04)	1.20(0.02) 1.96(0.03)	0.46(0.32) 0.10(0.90)	1.06(0.08) 2.71(0.00)	1.25(0.02) 2.26(0.02)	1.27(0.03) 1.56(0.10)	1.50(0.00) 2.53(0.00)	1.29(0.01) 2.80(0.00)	1.71(0.00) 2.18(0.04)
212	1.52(0.06) n = 108	1.37(0.01) n = 108	0.79(0.12) n = 108	**-3.83(0.00)** n = 104	1.01(0.15) n = 101	3.24(0.00) n = 107		0.14(0.79) 0.37(0.69)	0.61(0.19) 0.30(0.76)	0.40(0.45) 0.77(0.41)	0.31(0.49) 0.12(0.89)	0.67(0.20) 1.16(0.26)	1.04(0.06) 2.69(0.01)	0.81(0.09) 1.10(0.24)	0.82(0.13) 1.38(0.16)	1.55(0.01) 2.34(0.04)	0.96(0.06) 2.13(0.03)	1.18(0.02) 2.10(0.03)	0.57(0.26) 0.26(0.76)	1.14(0.06) 3.09(0.01)	1.29(0.02) 2.55(0.02)	0.96(0.03) 1.17(0.16)	1.55(0.01) 2.92(0.01)	1.13(0.02) 2.82(0.00)	1.36(0.01) 1.61(0.10)
4	**-1.91(0.03)** n = 107	**-2.08(0.01)** n = 107	**-2.67(0.00)** n = 107	**-7.18(0.00)** n = 103	**-2.57(0.00)** n = 100	-0.17(0.80) n = 106	**-3.50(0.00)** n = 107		0.39(0.41) -0.16(0.88)	0.36(0.40) 0.51(0.48)	0.31(0.56) 0.15(0.89)	0.49(0.25) 0.75(0.37)	0.97(0.10) 2.40(0.01)	0.74(0.15) 0.76(0.45)	0.72(0.13) 1.29(0.18)	1.46(0.01) 2.01(0.08)	0.82(0.11) 1.77(0.07)	0.84(0.10) 1.42(0.13)	0.50(0.26) -0.36(0.64)	1.08(0.06) 2.81(0.01)	1.16(0.03) 2.19(0.02)	1.14(0.02) 1.44(0.13)	1.44(0.01) 2.57(0.01)	1.22(0.01) 2.81(0.00)	1.64(0.00) 2.02(0.06)
298	3.10(0.00) n = 107	3.01(0.00) n = 107	2.38(0.00) n = 107	**-2.12(0.00)** n = 104	2.43(0.00) n = 100	4.89(0.00) n = 106	1.62(0.01) n = 107	5.14(0.00) n = 106		-0.13(0.77) 0.56(0.47)	-0.00(0.99) 0.63(0.52)	0.04(0.92) 0.83(0.23)	0.44(0.41) 2.45(0.01)	0.23(0.62) 0.76(0.37)	0.38(0.40) 1.56(0.08)	0.95(0.05) 2.06(0.03)	0.35(0.44) 1.74(0.03)	0.64(0.23) 2.18(0.03)	0.17(0.71) 0.29(0.72)	0.55(0.32) 2.85(0.00)	0.70(0.13) 2.27(0.02)	0.80(0.14) 1.96(0.07)	0.96(0.04) 2.64(0.00)	0.73(0.13) 2.92(0.00)	1.00(0.07) 1.97(0.06)
46	-0.27(0.70) n = 107	-0.47(0.57) n = 107	-1.09(0.19) n = 107	**-5.50(0.00)** n = 103	-0.65(0.45) n = 100	1.56(0.03) n = 106	**-1.90(0.02)** n = 107	1.71(0.05) n = 106	**-3.46(0.00)** n = 106		0.09(0.86) -0.05(0.96)	0.27(0.57) 0.32(0.71)	0.64(0.23) 1.95(0.03)	0.36(0.48) 0.25(0.80)	0.50(0.33) 0.97(0.33)	1.09(0.04) 1.49(0.18)	0.52(0.26) 1.33(0.13)	0.75(0.06) 1.35(0.10)	0.36(0.43) -0.31(0.68)	0.74(0.15) 2.35(0.01)	0.85(0.07) 1.74(0.08)	0.86(0.07) 1.26(0.18)	1.11(0.03) 2.13(0.03)	0.87(0.08) 2.29(0.01)	1.48(0.01) 1.91(0.08)
383	2.83(0.00) n = 103	2.68(0.00) n = 103	2.07(0.01) n = 103	**-2.58(0.00)** n = 101	2.33(0.01) n = 96	4.62(0.00) n = 102	1.23(0.06) n = 103	4.79(0.00) n = 102	-0.31(0.62) n = 103	3.37(0.00) n = 102		0.08(0.87) 0.31(0.70)	0.48(0.43) 2.20(0.03)	0.30(0.50) 0.52(0.49)	0.24(0.59) 0.80(0.28)	1.04(0.04) 1.67(0.07)	0.35(0.51) 1.28(0.14)	0.61(0.18) 1.42(0.07)	0.30(0.57) 0.19(0.81)	0.52(0.36) 2.34(0.01)	0.68(0.18) 1.73(0.06)	0.60(0.02) 0.94(0.03)	0.96(0.07) 2.12(0.02)	0.71(0.15) 2.35(0.00)	1.05(0.06) 1.40(0.13)
351	4.73(0.00) n = 106	4.59(0.00) n = 106	3.96(0.00) n = 106	-0.44(0.39) n = 103	4.07(0.00) n = 99	6.45(0.00) n = 105	3.13(0.00) n = 106	6.78(0.00) n = 105	1.60(0.00) n = 106	4.87(0.00) n = 105	1.89(0.00) n = 102		0.37(0.50) 1.56(0.10)	0.16(0.72) -0.11(0.88)	0.26(0.51) 0.64(0.37)	0.85(0.09) 1.17(0.18)	0.32(0.49) 1.01(0.18)	0.58(0.30) 1.11(0.21)	0.08(0.87) -0.55(0.48)	0.50(0.37) 1.99(0.02)	0.68(0.18) 1.43(0.08)	0.72(0.16) 1.02(0.24)	0.94(0.04) 1.85(0.02)	0.69(0.12) 2.02(0.01)	1.07(0.04) 1.18(0.21)
137	2.19(0.02) n = 107	2.01(0.01) n = 107	1.40(0.06) n = 107	**-3.15(0.00)** n = 103	1.56(0.03) n = 100	3.85(0.00) n = 106	0.64(0.35) n = 107	4.24(0.00) n = 106	-1.02(0.13) n = 106	2.59(0.00) n = 106	-0.89(0.31) n = 102	**-2.60(0.00)** n = 105		-0.24(0.63) -1.57(0.08)	-0.04(0.94) -0.97(0.27)	0.55(0.27) -0.23(0.81)	-0.07(0.89) -0.53(0.54)	0.19(0.74) -0.55(0.58)	0.45(0.40) -1.09(0.21)	0.11(0.68) 0.46(0.43)	0.31(0.56) -0.01(0.99)	0.35(0.52) -0.66(0.48)	0.55(0.20) 0.34(0.69)	0.30(0.52) 0.59(0.43)	0.87(0.09) -0.03(0.98)
267	1.73(0.04) n = 107	1.50(0.05) n = 107	0.96(0.22) n = 107	**-3.54(0.00)** n = 103	1.30(0.11) n = 100	3.51(0.00) n = 106	0.20(0.76) n = 107	3.70(0.00) n = 106	**-1.43(0.01)** n = 106	2.02(0.01) n = 106	-1.16(0.08) n = 102	**-3.01(0.00)** n = 105	-0.55(0.38) n = 106		0.16(0.70) 0.68(0.33)	0.72(0.09) 1.21(0.11)	0.08(0.84) 0.96(0.17)	0.37(0.42) 1.16(0.19)	-0.06(0.90) -0.48(0.57)	0.34(0.51) 2.04(0.01)	0.45(0.30) 1.41(0.07)	0.52(0.25) 0.81(0.34)	0.72(0.09) 1.78(0.01)	0.46(0.26) 1.85(0.02)	0.84(0.05) 1.07(0.16)
349	2.97(0.00) n = 105	2.97(0.00) n = 105	2.30(0.00) n = 105	**-2.09(0.00)** n = 101	2.67(0.00) n = 99	4.65(0.00) n = 104	1.55(0.02) n = 105	4.91(0.00) n = 104	0.02(0.98) n = 104	3.41(0.00) n = 104	0.11(0.87) n = 101	**-1.66(0.00)** n = 103	0.99(0.13) n = 104	1.40(0.01) n = 104		0.60(0.18) 0.66(0.44)	-0.02(0.96) 0.30(0.68)	0.25(0.64) 0.45(0.60)	-0.01(0.99) -0.67(0.39)	0.18(0.73) 1.33(0.08)	0.33(0.52) 0.77(0.36)	0.44(0.35) 0.31(0.71)	0.59(0.14) 1.15(0.09)	0.39(0.26) 1.48(0.02)	0.58(0.16) 0.24(0.75)
248	1.41(0.12) n = 108	1.26(0.14) n = 108	0.68(0.43) n = 108	**-3.75(0.00)** n = 104	1.07(0.21) n = 101	3.18(0.00) n = 107	-0.11(0.89) n = 108	3.40(0.00) n = 107	**-1.74(0.02)** n = 107	1.72(0.05) n = 107	-1.50(0.07) n = 103	**-3.31(0.00)** n = 106	-0.72(0.31) n = 107	-0.31(0.62) n = 107	**-1.72(0.01)** n = 105		-0.64(0.20) -0.22(0.79)	-0.46(0.38) -0.27(0.78)	-0.44(0.34) -0.97(0.28)	-0.40(0.42) 0.75(0.38)	-0.25(0.62) 0.21(0.82)	-0.40(0.44) -0.59(0.55)	0.01(0.99) 0.58(0.46)	-0.22(0.63) 0.78(0.37)	0.11(0.81) -0.04(0.96)
414	1.93(0.01) n = 107	1.84(0.01) n = 107	1.25(0.06) n = 107	**-3.37(0.00)** n = 104	1.46(0.04) n = 100	3.61(0.00) n = 106	0.45(0.49) n = 107	3.86(0.00) n = 106	**-1.24(0.04)** n = 106	2.20(0.00) n = 106	-0.97(0.18) n = 102	**-2.85(0.00)** n = 105	-0.16(0.81) n = 106	0.18(0.77) n = 106	-1.13(0.06) n = 104	0.45(0.54) n = 107		0.34(0.41) 0.18(0.83)	0.04(0.93) -0.97(0.14)	0.20(0.69) 1.01(0.19)	0.35(0.28) 0.44(0.42)	0.41(0.44) 0.01(0.99)	0.61(0.02) 0.81(0.04)	0.29(0.46) 0.83(0.19)	0.73(0.08) -0.02(0.98)
609	2.80(0.00) n = 99	2.33(0.00) n = 99	1.80(0.01) n = 99	**-2.62(0.00)** n = 96	2.07(0.01) n = 92	4.34(0.00) n = 98	0.92(0.11) n = 99	4.17(0.00) n = 99	-0.65(0.32) n = 99	2.71(0.00) n = 98	-0.54(0.47) n = 96	**-1.97(0.00)** n = 98	0.42(0.56) n = 98	0.68(0.28) n = 98	-0.53(0.43) n = 96	0.84(0.23) n = 99	0.69(0.25) n = 98		-0.22(0.68) -1.04(0.21)	-0.10(0.85) 0.92(0.28)	0.25(0.49) 0.82(0.28)	0.09(0.84) -0.33(0.64)	0.31(0.50) 0.61(0.48)	0.08(0.90) 0.91(0.24)	0.47(0.38) 0.16(0.84)
275	5.00(0.00) n = 90	5.23(0.00) n = 90	4.81(0.00) n = 90	0.36(0.40) n = 86	5.20(0.00) n = 86	6.76(0.00) n = 89	3.92(0.00) n = 90	7.10(0.00) n = 89	2.01(0.00) n = 89	5.58(0.00) n = 89	2.19(0.00) n = 87	0.66(0.17) n = 88	4.05(0.00) n = 89	3.75(0.00) n = 89	2.81(0.00) n = 90	4.59(0.00) n = 90	3.79(0.00) n = 89	2.98(0.00) n = 84		-0.41(0.42) 1.19(0.13)	0.29(0.58) 1.39(0.10)	0.16(0.75) 0.52(0.47)	0.46(0.29) 1.64(0.03)	0.06(0.83) 1.64(0.00)	0.74(0.01) 0.98(0.10)
136	2.20(0.02) n = 108	2.05(0.00) n = 108	1.46(0.03) n = 108	**-3.04(0.00)** n = 104	1.56(0.02) n = 101	3.95(0.00) n = 107	0.67(0.32) n = 108	4.23(0.00) n = 107	-0.98(0.16) n = 107	2.59(0.00) n = 107	-0.87(0.31) n = 103	**-2.52(0.00)** n = 106	0.03(0.93) n = 107	0.54(0.41) n = 107	-0.98(0.11) n = 105	0.78(0.26) n = 108	0.28(0.66) n = 107	-0.43(0.50) n = 99	**-4.19(0.00)** n = 90		0.15(0.77) -0.54(0.52)	0.27(0.59) -1.07(0.17)	0.41(0.34) -0.17(0.80)	0.19(0.66) 0.02(0.98)	0.82(0.10) -0.47(0.52)
415	2.22(0.01) n = 108	2.07(0.01) n = 108	1.49(0.04) n = 108	**-3.07(0.00)** n = 104	1.63(0.04) n = 101	3.95(0.00) n = 107	0.70(0.26) n = 108	4.19(0.00) n = 107	-0.94(0.08) n = 107	2.51(0.00) n = 107	-0.83(0.23) n = 103	**-2.46(0.00)** n = 106	0.09(0.88) n = 107	0.48(0.40) n = 107	-0.92(0.14) n = 105	0.81(0.26) n = 108	0.31(0.42) n = 107	-0.26(0.63) n = 99	**-3.64(0.00)** n = 90	0.02(0.97) n = 108		0.11(0.84) -0.45(0.63)	0.26(0.47) 0.37(0.56)	-0.07(0.87) 0.43(0.56)	0.04(0.92) -1.08(0.12)
243	4.09(0.00) n = 101	3.84(0.00) n = 101	3.28(0.00) n = 101	**-1.49(0.02)** n = 97	3.71(0.00) n = 96	5.68(0.00) n = 100	2.39(0.00) n = 101	5.76(0.00) n = 100	0.91(0.13) n = 100	4.40(0.00) n = 100	0.88(0.02) n = 99	-0.68(0.20) n = 99	2.04(0.01) n = 100	2.37(0.00) n = 100	0.96(0.13) n = 100	2.39(0.00) n = 101	2.14(0.00) n = 100	1.67(0.02) n = 93	**-1.72(0.01)** n = 88	2.09(0.01) n = 101	1.96(0.00) n = 101		0.24(0.64) 1.02(0.26)	-0.05(0.91) 1.13(0.11)	0.50(0.35) 0.71(0.42)
413	2.58(0.00) n = 108	2.42(0.00) n = 108	1.84(0.01) n = 108	**-2.66(0.00)** n = 104	2.03(0.01) n = 101	4.32(0.00) n = 107	1.05(0.11) n = 108	4.57(0.00) n = 107	-0.58(0.34) n = 107	2.91(0.00) n = 107	-0.39(0.59) n = 103	**-2.14(0.00)** n = 106	0.46(0.42) n = 107	0.83(0.15) n = 107	-0.54(0.33) n = 105	1.16(0.10) n = 108	0.61(0.07) n = 107	-0.05(0.94) n = 99	**-3.31(0.00)** n = 90	0.38(0.49) n = 108	0.36(0.40) n = 108	**-1.52(0.03)** n = 101		-0.28(0.45) 0.02(0.98)	0.03(0.94) -0.82(0.20)
347	4.27(0.00) n = 105	3.99(0.00) n = 105	3.42(0.00) n = 105	-0.91(0.06) n = 101	4.09(0.00) n = 98	6.16(0.00) n = 104	2.51(0.00) n = 105	6.39(0.00) n = 104	1.00(0.06) n = 104	4.68(0.00) n = 104	1.21(0.06) n = 101	-0.40(0.43) n = 103	2.12(0.00) n = 105	2.55(0.00) n = 104	1.33(0.01) n = 102	2.98(0.00) n = 105	2.43(0.00) n = 104	1.74(0.01) n = 97	**-1.53(0.00)** n = 88	2.17(0.00) n = 105	2.02(0.00) n = 105	0.34(0.57) n = 99	1.61(0.00) n = 105		0.46(0.11) -0.70(0.16)
274	5.64(0.00) n = 96	5.28(0.00) n = 96	4.78(0.00) n = 96	0.47(0.33) n = 92	5.30(0.00) n = 90	7.45(0.00) n = 96	3.99(0.00) n = 96	7.61(0.00) n = 95	2.48(0.00) n = 95	6.16(0.00) n = 95	2.81(0.00) n = 92	1.01(0.06) n = 94	3.71(0.00) n = 96	4.02(0.00) n = 96	2.60(0.00) n = 94	4.10(0.00) n = 96	3.86(0.00) n = 95	3.31(0.00) n = 90	0.13(0.75) n = 82	3.74(0.00) n = 96	3.34(0.00) n = 96	1.79(0.00) n = 91	3.09(0.00) n = 96	1.51(0.00) n = 95	

The results of paired Student's *t*-test on common targets are reported in the form of the mean of the differences in GDT-HA, AL-0, and HBscore values along with the associated P-values in parentheses. Cells above the diagonal in the top right part of the table provide values for GDT-HA (upper half of each cell) and AL0 (lower half of each cell) comparing rows with columns. Cells below the diagonal in the lower left part of the table provide values for HBscore (upper half of each cell) comparing columns with rows, and the number of common models in the pair wise comparison (lower half of each cell). Statistically significant differences between groups (P-values < 0.05) are shaded in gray.

DOI 10.1002/prot

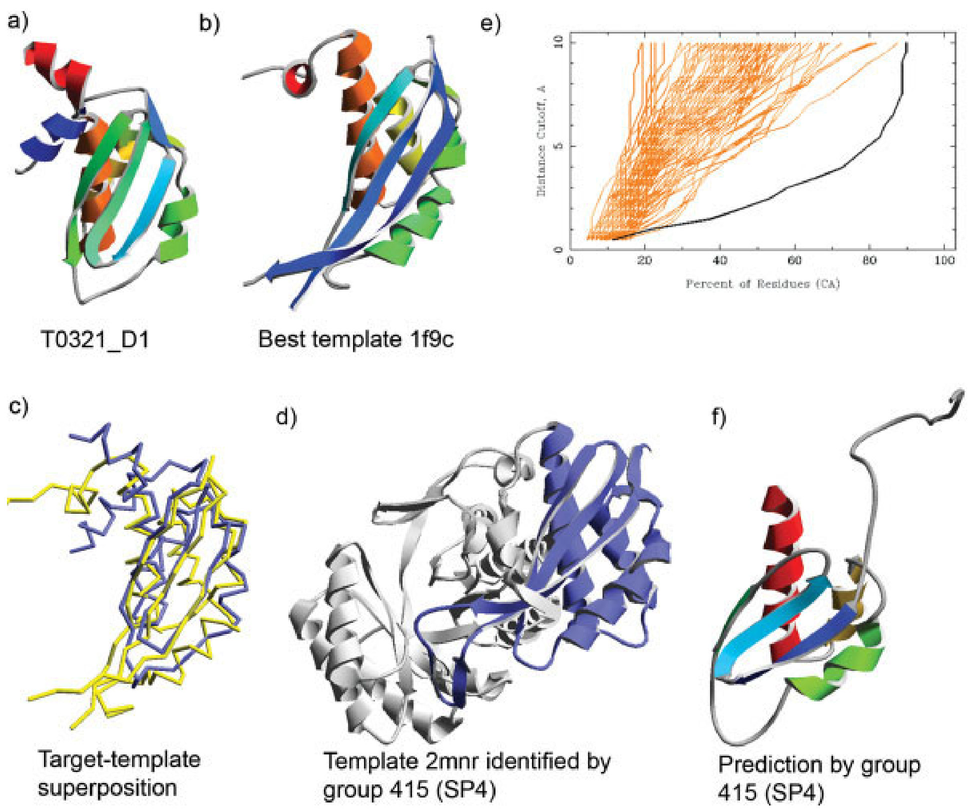

a)
b)
e)
T0321_D1
Best template 1f9c
Distance Cutoff, A
Percent of Residues (CA)
0
5
10
0
20
40
60
80
100
c)
d)
f)
Target-template superposition
Template 2mnr identified by group 415 (SP4)
Prediction by group 415 (SP4)

Figure 2

Prediction example of target T0321 domain 1. For discussion, see the main document.

was able to improve over the PSI-BLAST template by a margin of over 10 GDT-HA points in 46 cases, and by a margin of more than 20 points in 23 cases. It was also able to build models of comparable quality to the "LGA-predictor" in the majority of cases. Only in 23 cases, the server failed to get within 10 GDT-HA points of the LGA-based pseudopredictor, and the predictions were more than 20 GDT-HA points lower in eight cases.

As is apparent from Figure 4 (lower panel), the performance of the PSI-BLAST approach is only weakly correlated with the overall structural similarity of the best available template (correlation coefficient [GDT-HA(LGA-pseu-

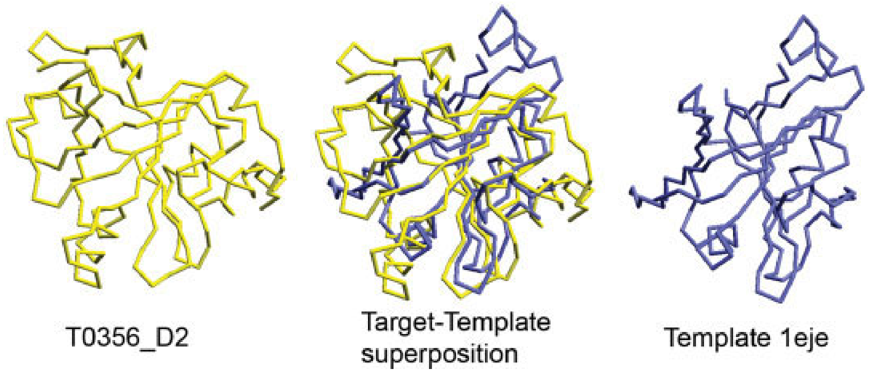


Figure 3

Prediction example of target T0356 domain 2. For discussion, see main document.

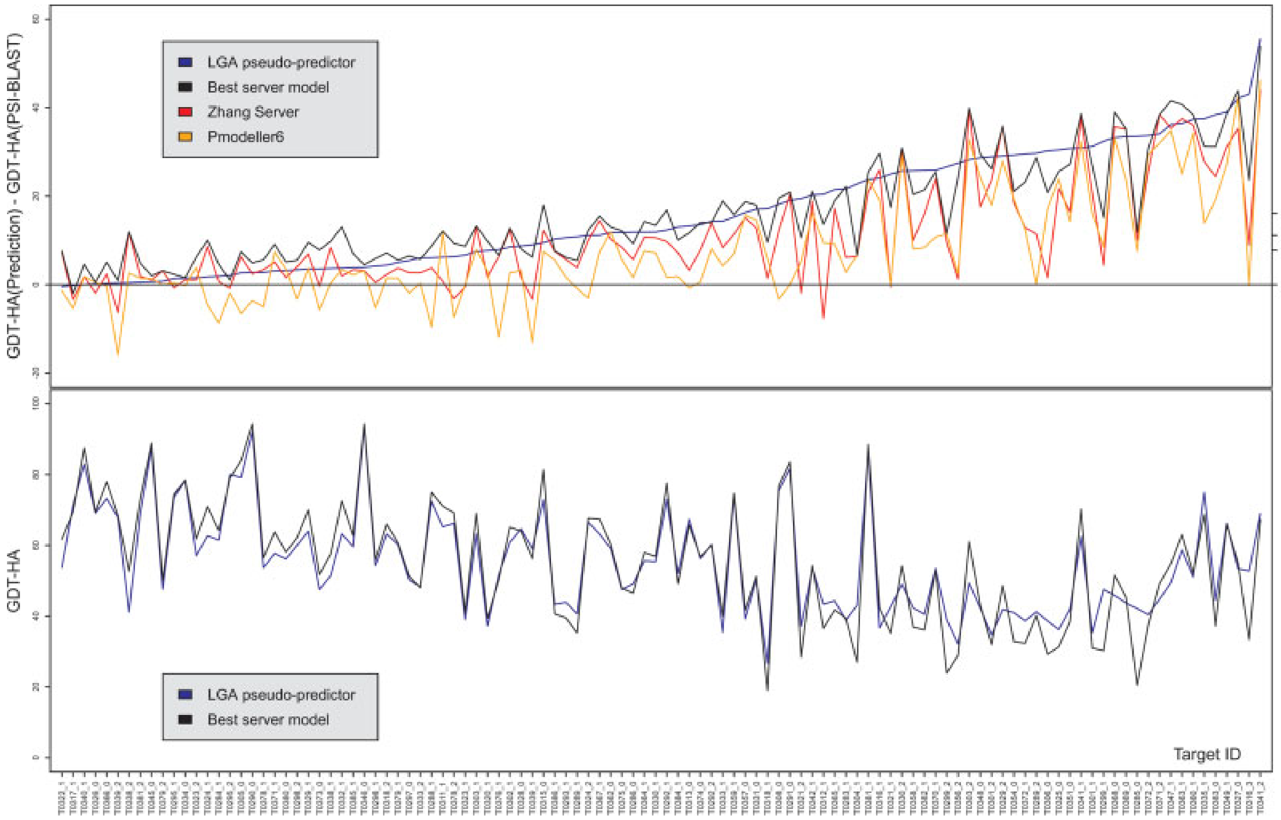


Figure 4

Server performance compared with the two pseudopredictors. Above: For each target, the performance of the different predictors is plotted relative to the GDT-HA baseline defined by the "PSI-BLAST" pseudopredictor, i.e., values on the vertical axis are GDT-HA values minus the GDT-HA of the pseudoprediction based on PSI-BLAST. Points below the x-axis performed worse than a naïve PSI-BLAST alignment would have fared. Targets are ordered by increasing difference in GDT-HA between pseudopredictions based on templates identified by LGA and PSI-BLAST, respectively. Predictions by group 25 (Zhang-server) are shown in red, 47 (Pmodeller6) in orange, and the "LGA pseudopredictor" in blue. The best of all models submitted by any server group are shown in black. For comparison, the average performance of the methods is indicated on the right side of the plot: (A) PSI-BLAST pseudo predictor, (B) Pmodeller6, (C) Zhang-server, and (D) LGA-pseudopredictor. Below: For reference, the absolute GDT-HA value is plotted for the "LGA pseudopredictor" (blue) and the best server submission per target in retrospect (black).

domodel) − GDT-HA(PSI-BLAST-pseudomodel] versus GDT-HA(LGA-pseudomodel) = −0.42). There were numerous cases where good structural templates were available (as identified by LGA), but were missed by PSI-BLAST.

Some cases, in which good templates existed but could not be detected easily, have been identified and rationalized. One example of such a notable improvement over the PSI-BLAST performance is T0349_D1, a NMR structure of the hypothetical protein RPA1041 from *Pseudomonas aeruginosa*. A model based on the best available template, a secretory protein of the YscJ/FliF family part of the *E. coli* type III secretion system (PDB:1yj7, chain D), has a GDT-HA score of 66.2. The template detected by PSI-BLAST, the structure of the L-aminopeptidase D-ala-esterase/amidase from *Ochrobactrum anthropi* (PDB: 1b65), has a high target coverage of 72%, but an unfavorable e-value and is structurally unrelated. Consequently, the resulting model achieves a GDT-HA value of only 27.2. The reason for PSI-BLAST's inability to detect a better template may be due to the meager yield of hits during the profile creation step, which consequently leads to a poor profile for further scanning. It is noteworthy that some of the better performing servers, e.g., HHpred1, work by matching profiles generated from the target and the potential templates. Presumably, this two-sided approach allows the sequence gap to be bridged successfully where the one-sided PSI-BLAST method fails.

SERVER PREDICTIONS IN THE HA CATEGORY

One might expect that servers will do well when evolutionary relationships and sequence alignments are clear, as is generally the case for the structures in the HA/TBM category. To enter this category, it is necessary that there be a good template, and in most cases, such templates were successfully identified. Indeed, the trend for servers to perform well continued in this category.

To obtain a rough overall ranking of the groups submitting predictions for the HA/TBM category, they were sorted by the sum of the scores for GDT-HA, prediction of side-chain χ_1/χ_2 angles, and suitability for use as models to solve the target crystal structures by molecular replacement.[68] Of the top 25 groups, four are servers. Two of the servers are among the top five, with similar overall scores placing them at positions 4 (186 CaspItaFOX) and 5 (4 Robetta). However, in contrast to the results for the general TBM category, Zhang-server did not appear among the top groups, coming in at position 43. This ranking is strongly influenced by the performance in rotamer prediction and molecular replacement, where CaspItaFOX and Robetta were both highly ranked. A number of servers did extremely well judged by the more traditional GDT-HA score, but less well against the other two criteria. In fact, when the predictors are ranked by GDT-HA alone, servers are found at positions 3 (136 FOLDpro), 4 (25 Zhang-server), and 5 (137 3Dpro).

SERVER PREDICTIONS IN THE FM CATEGORY

Servers might be expected to be at a greater disadvantage in FM than in TBM. For one thing, human insight can still be useful in modeling protein structures, and it would seem that there are more opportunities for such intervention in FM than in TBM. Second, the time limit for server predictions might be more of a constraint for FM targets than for TBM. Nevertheless, servers did rather well on FM targets in CASP7.

Although the best group in FM was clearly a human group (20 Baker), the next set of groups includes servers. On the basis of the pairwise GDT-TS comparison, the Zhang-server (Group 25) was the third best FM predictor, behind only the human-aided predictions of Baker and Zhang himself (Group 24). By visual assessment, two servers (25 Zhang-server and 4 Robetta) were in a four-way tie for second place according to the criteria and scoring scheme applied in the FM category of CASP7.[66] Combining GDT-TS assessment and visual assessment, there were six groups that were among the top 20 by GDT-TS and among the top 10 by visual assessment. Among these six groups, three were servers (Zhang-server, Robetta, and the metaserver group 47 Pmodeller6).

A particularly striking server prediction was the Robetta model 2 for target T0350, which has an antiparallel three-stranded sheet and three flanking helices that lie side by side against one surface of the sheet. Although a relatively simple αβ-structure, the arrangement of the sheet and helices is unusual and appears to be a true FM target. Nevertheless, many groups did astonishingly well on this target, and of the models that were judged best by visual assessment, Model 2 from Robetta had the highest GDT-TS rank. For two other targets (T0287 and T0314), the Zhang-server had the single best model by GDT-TS.

DISCUSSION

In the recent years, structure prediction in CASP had been dominated by human predictors using computational modeling procedures, which require manual intervention at many steps in the process. One of the recurring problems in CASP assessments has always been discerning the improvements contributed by human intervention from the purely computational element. In the assessment of CASP2, Thornton and coworkers used the automated modeling service SWISS-MODEL as reference to establish a baseline of what could be achieved by fully automated prediction.[78,79] During the following experiments, numerous predictors registered their methods as servers and were assessed separately as part of the CAFASP experiment series.[8,9] Starting with CASP6, server predictions were assessed as part of the main CASP experiment.[10] The general opinion in the community has been that the “human plus machine predictions” are superior to automated ones.[8] However, it appears that with CASP7, this view might have to be revised as 6 of the top 25 groups in the TBM category assessment were server predictors. Overall, in both CASP5 and CASP6 servers provided the best models (or tied with humans) for 7% of targets as measured by GDT-TS.[7] In contrast, for the 123 target domains in CASP7, the best model (or tied with humans) was submitted by a server in 29% of the cases. The best prediction server (25 Zhang-server) was ranked third over all, i.e. it outperformed all but two of the participating groups in the TBM category.

With the emergence of many new protein structure prediction servers, based on diverse methods with different strengths and weaknesses, the question of selecting the most appropriate server for each target became prevailing. Metaservers—methods that use the results of other servers as input to generate their predictions—were expected to have the capacity to outperform all individual autonomous servers, and challenge most human expert predictors.[80] To illustrate to what extent metaservers were able to select favorable models from the pool of the CASP7 server models, we included in Figure 4 the best-performing metaserver (47 Pmodeller6), the best individual server (25 Zhang-server), and the best of all submitted server models. Remarkably in CASP7, the best individual autonomous server (25 Zhang-server) outperformed the best metaserver (47 Pmodeller6), as well as the best manual metapredictor group (675 Famsace).[67] Comparing all server methods among themselves (Fig. 1), the top ranking groups are not metamethods, indicating that the development of individual servers was fruitful in the recent years, and has significantly contributed to advancing the field of protein structure prediction.

There has been much debate on how to objectively compare human and server predictions. Much effort is expended during CASP by human predictors to gather information from a diverse set of resources such as scientific literature and specialist databases. This information may contribute greatly to the quality of a model for example by improving template selection and alignment or identifying ligand-binding sites. As this work, however, is very time consuming, it is only feasible for a relatively small set of targets. By contrast, prediction servers are fully automated and do not have these limitations, therefore their performance can be evaluated using a larger sample size than would be possible with human predictors. Several projects have been initiated with the aim of continuous, large-scale assessment, such as LiveBench[80] and EVA,[81] which use a sample size not tractable for nonautomated prediction methods. The CASP7 experiment comprised the relatively large number of 100 prediction targets, providing a solid basis for the numerical and statistical analysis. In this respect, CASP7 reflects a "real life" situation, where one is faced with the problem of modeling structures for an exponentially growing number of protein sequences. Although the large number of targets led to a technical advantage for automated methods over human predictors, the server predictions for all targets were publicly available early in the prediction window. Thus, most of the "routine" work had already been done automatically and human predictors could focus their efforts on improving the automated predictions using expert knowledge.

The gap between human predictors and servers is closing as automated prediction servers have come of age. However, the fundamental limitations for both human and server predictors remain in modeling of loops and effective refinement techniques. The observed progress in automated, reproducible, and scalable prediction methods in CASP7 holds the promise for further improvements in the future.

ACKNOWLEDGMENTS

We thank the team at the Prediction Center at UC Davis, especially Andriy Kryshtafovych, for their professional support, and Michael Tress for providing LGA template information. We also thank Anna Tramontano, Alfonso Valencia, Burkhard Rost, Tim Hubbard, and John Moult for fruitful and encouraging discussions. We thank all server groups who participated in the survey for this manuscript, and last but not least, we acknowledge all participating experimental groups, CASP7 predictors, and server developers, without whom CASP would not be possible.

REFERENCES

1. Friedberg I, Jaroszewski L, Ye Y, Godzik A. The interplay of fold recognition and experimental structure determination in structural genomics. Curr Opin Struct Biol 2004;14:307–312.
2. Marsden RL, Lewis TA, Orengo CA. Towards a comprehensive structural coverage of completed genomes: a structural genomics viewpoint. BMC Bioinformatics 2007;8:86.
3. Fox JA, McMillan S, Ouellette BFF. A compilation of molecular biology web servers: 2006 update on the Bioinformatics Links Directory. Nucleic Acids Res 2006;34(Suppl 2):W3–W5.
4. Pieper U, Eswar N, Braberg H, Madhusudhan MS, Davis FP, Stuart AC, Mirkovic N, Rossi A, Marti-Renom MA, Fiser A, Webb B, Greenblatt D, Huang CC, Ferrin TE, Sali A. MODBASE, a database of annotated comparative protein structure models, and associated resources. Nucleic Acids Res 2004;32(Suppl 1):D217–D222.
5. Kopp J, Schwede T. The SWISS-MODEL repository: new features and functionalities. Nucleic Acids Res 2006;34(Suppl 1):D315–D318.
6. Castrignano T, De Meo PD, Cozzetto D, Talamo IG, Tramontano A. The PMDB protein model database. Nucleic Acids Res 2006;34 (Database Issue):D306–D309.
7. Moult J. A decade of CASP: progress, bottlenecks and prognosis in protein structure prediction. Curr Opin Struct Biol 2005;15:285–289.
8. Fischer D, Barret C, Bryson K, Elofsson A, Godzik A, Jones D, Karplus KJ, Kelley LA, MacCallum RM, Pawowski K, Rost B, Rychlewski L, Sternberg M. CAFASP-1: critical assessment of fully automated structure prediction methods. Proteins 1999;(Suppl 3):209–217.
9. Fischer D, Rychlewski L, Dunbrack RL, Jr, Ortiz AR, Elofsson A. CAFASP3: the third critical assessment of fully automated structure prediction methods. Proteins 2003;53(Suppl 6):503–516.
10. Moult J, Fidelis K, Rost B, Hubbard T, Tramontano A. Critical assessment of methods of protein structure prediction (CASP)–round 6. Proteins 2005;61(Suppl 7):3–7.
11. Kryshtafovych A, Venclovas C, Fidelis K, Moult J. Progress over the first decade of CASP experiments. Proteins 2005;61(Suppl 7):225–236.
12. Bates PA, Kelley LA, MacCallum RM, Sternberg MJ. Enhancement of protein modeling by human intervention in applying the automatic programs 3D-JIGSAW and 3D-PSSM. Proteins 2001;45(Suppl 5):39–46.
13. Offman MN, Fitzjohn PW, Bates PA. Developing a move-set for protein model refinement. Bioinformatics 2006;22:1838–1845.
14. Contreras-Moreira B, Fitzjohn PW, Offman M, Smith GR, Bates PA. Novel use of a genetic algorithm for protein structure prediction: searching template and sequence alignment space. Proteins 2003;53(Suppl 6):424–429.
15. Cheng J, Randall AZ, Sweredoski MJ, Baldi P. SCRATCH: a protein structure and structural feature prediction server. Nucleic Acids Res 2005;33(Web Server Issue):W72–W76.
16. Kim DE, Chivian D, Malmstrom L, Baker D. Automated prediction of domain boundaries in CASP6 targets using Ginzu and RosettaDOM. Proteins 2005;61(Suppl 7):193–200.
17. Cheng J, Baldi P. Three-stage prediction of protein β-sheets by neural networks, alignments and graph algorithms. Bioinformatics 2005;21(Suppl 1):i75–i84.
18. Ishida T, Nishimura T, Nozaki M, Inoue T, Terada T, Nakamura S, Shimizu K. Development of an ab initio protein structure prediction system ABLE. Genome Inform 2003;14:228–237.
19. Su CT, Chen CY, Ou YY. Protein disorder prediction by condensed PSSM considering propensity for order or disorder. BMC Bioinformatics 2006;7:319.
20. Lund O, Nielsen M, Lundegaard C, Worning P. CPHmodels 2.0: X3M a Computer Program to Extract 3D Models. In: CASP5: Proceedings of the 5th meeting on the critical assessment of techniques for protein structure prediction, 1–5 December 2002, Asilomar, CA.

DOI 10.1002/prot

21. Ward JJ, McGuffin LJ, Bryson K, Buxton BF, Jones DT. The DISOPRED server for the prediction of protein disorder. Bioinformatics 2004;20:2138–2139.
22. Cheng J, Sweredoski MJ, Baldi P. Accurate prediction of protein disordered regions by mining protein structure data. Data Min Knowledge Discov 2005;11:213–222.
23. Bau D, Martin AJ, Mooney C, Vullo A, Walsh I, Pollastri G. Distill: a suite of web servers for the prediction of one-, two- and three-dimensional structural features of proteins. BMC Bioinformatics 2006;7:402.
24. Bryson K, Cozzetto D, Jones DT. Computer-assisted protein domain boundary prediction using the DomPred server. Curr Protein Pept Sci 2007;8:181–188.
25. Ogata K, Umeyama H. An automatic homology modeling method consisting of database searches and simulated annealing. J Mol Graph Model 2000;18:258–272, 305–256.
26. Cheng J, Baldi P. A machine learning information retrieval approach to protein fold recognition. Bioinformatics 2006;22:1456–1463.
27. Tomii K, Akiyama Y. FORTE: a profile-profile comparison tool for protein fold recognition. Bioinformatics 2004;20:594–595.
28. Tomii K, Hirokawa T, Motono C. Protein structure prediction using a variety of profile libraries and 3D verification. Proteins 2005;61 (Suppl 7):114–121.
29. Kosinski J, Gajda MJ, Cymerman IA, Kurowski MA, Pawlowski M, Boniecki M, Obarska A, Papaj G, Sroczynska-Obuchowicz P, Tkaczuk KL, Sniezynska P, Sasin JM, Augustyn A, Bujnicki JM, Feder M. FRankenstein becomes a cyborg: the automatic recombination and realignment of fold recognition models in CASP6. Proteins 2005;61(Suppl 7):106–113.
30. Shi J, Blundell TL, Mizuguchi K. FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. J Mol Biol 2001;310: 243–257.
31. Kurowski MA, Bujnicki JM. GeneSilico protein structure prediction meta-server. Nucleic Acids Res 2003;31:3305–3307.
32. MacCallum RM. Striped sheets and protein contact prediction. Bioinformatics 2004;20(Suppl 1):I224–I231.
33. Soding J, Biegert A, Lupas AN. The HHpred interactive server for protein homology detection and structure prediction. Nucleic Acids Res 2005;33(Web Server Issue):W244–W248.
34. Torda AE, Procter JB, Huber T. Wurst: a protein threading server with a structural scoring function, sequence profiles and optimized substitution matrices. Nucleic Acids Res 2004;32 (Web Server Issue):W532–W535.
35. Rangwala H, Karypis G. Profile-based direct kernels for remote homology detection and fold recognition. Bioinformatics 2005;21: 4239–4247.
36. Rangwala H, Karypis G. Building multiclass classifiers for remote homology detection and fold recognition. BMC Bioinformatics 2006;7:455.
37. Kalisman N, Levi A, Maximova T, Reshef D, Zafriri-Lynn S, Gleyzer Y, Keasar C. MESHI: a new library of Java classes for molecular modeling. Bioinformatics 2005;21:3931–3932.
38. Meller J, Elber R. Linear programming optimization and a double statistical filter for protein threading protocols. Proteins 2001;45: 241–261.
39. Teodorescu O, Galor T, Pillardy J, Elber R. Enriching the sequence substitution matrix by structural information. Proteins 2004;54:41–48.
40. Tobi D, Elber R. Distance-dependent, pair potential for protein folding: results from linear optimization. Proteins 2000;41:40–46.
41. Saini HK, Fischer D. Meta-DP: domain prediction meta-server. Bioinformatics 2005;21:2917–2920.
42. Pandit SB, Zhang Y, Skolnick J. TASSER-Lite: an automated tool for protein comparative modeling. Biophys J 2006;91:4180–4190.
43. Zhang Y, Skolnick J. Automated structure prediction of weakly homologous proteins on a genomic scale. Proc Natl Acad Sci USA 2004;101:7594–7599.
44. Zhou H, Pandit SB, Lee SY, Borreguero J, Chen H, Wroblewska L, Skolnick J. Analysis of TASSER based CASP7 protein structure prediction results. Proteins 2007;69(Suppl 8):90–97.
45. Jones DT, Bryson K, Coleman A, McGuffin LJ, Sadowski MI, Sodhi JS, Ward JJ. Prediction of novel and analogous folds using fragment assembly and fold recognition. Proteins 2005;61 (Suppl 7):143–151.
46. Wallner B, Larsson P, Elofsson A. Pcons.net: protein structure prediction meta server. Nucleic Acids Res 2007;35 (Web Server Issue):W369–W374.
47. Hawkins T, Luban S, Kihara D. Enhanced automated function prediction using distantly related sequences and contextual association by PFP. Protein Sci 2006;15:1550–1556.
48. Hamilton N, Burrage K, Ragan MA, Huber T. Protein contact prediction using patterns of correlation. Proteins 2004;56:679–684.
49. Punta M, Rost B. PROFcon: novel prediction of long-range contacts. Bioinformatics 2005;21:2960–2968.
50. Hung LH, Ngan SC, Liu T, Samudrala R. PROTINFO: new algorithms for enhanced protein structure predictions. Nucleic Acids Res 2005;33(Web Server Issue):W77–W80.
51. Raghava GPS. MANGO: prediction of genome ontology (GO) class of a protein from its amino acid and dipeptide composition using nearest neighbor approach. In: CASP7: Proceedings of the 7th meeting on the critical assessment of techniques for protein structure prediction, 26–30 November 2006, Asilomar, CA.
52. Xu J, Li M, Kim D, Xu Y. RAPTOR: optimal protein threading by linear programming. J Bioinform Comput Biol 2003;1:95–117.
53. Bu D, Li S, Gao X, Yu L, Xu J, Li M. Consensus approaches for protein structure prediction. In: Zhang YQ, Jagath CR, editors. Machine Learning in Bioinformatics. New York: Wiley; 2007.
54. Chivian D, Kim DE, Malmstrom L, Bradley P, Robertson T, Murphy P, Strauss CE, Bonneau R, Rohl CA, Baker D. Automated prediction of CASP-5 structures using the Robetta server. Proteins 2003;53 (Suppl 6):524–533.
55. Schlessinger A, Rost B. Protein flexibility and rigidity predicted from sequence. Proteins 2005;61:115–126.
56. Schlessinger A, Yachdav G, Rost B. PROFbval: predict flexible and rigid residues in proteins. Bioinformatics 2006;22:891–893.
57. Karplus K, Katzman S, Shackleford G, Koeva M, Draper J, Barnes B, Soriano M, Hughey R. SAM-T04: what is new in protein-structure prediction for CASP6. Proteins 2005;61(Suppl 7):135–142.
58. Karplus K, Karchin R, Draper J, Casper J, Mandel-Gutfreund Y, Diekhans M, Hughey R. Combining local-structure, fold-recognition, and new fold methods for protein structure prediction. Proteins 2003;53(Suppl 6):491–496.
59. Karplus K, Barrett C, Cline M, Diekhans M, Grate L, Hughey R. Predicting protein structure using only sequence information. Proteins 1999;37(Suppl 3):121–125.
60. Zhou H, Zhou Y. Fold recognition by combining sequence profiles derived from evolution and from depth-dependent structural alignment of fragments. Proteins 2005;58:321–328.
61. Liu S, Zhang C, Liang S, Zhou Y. Fold Recognition by concurrent use of solvent accessibility and residue depth. Proteins 2007;68:636–645.
62. Cheng J, Baldi P. Improved residue contact prediction using support vector machines and a large feature set. BMC Bioinformatics 2007;8:113.
63. Poleksic A, Danzer JF, Hambly K, Debe DA. Convergent island statistics: a fast method for determining local alignment score significance. Bioinformatics 2005;21:2827–2831.
64. Debe DA, Danzer JF, Goddard WA, Poleksic A. STRUCTFAST: protein sequence remote homology detection and alignment using novel dynamic programming and profile-profile scoring. Proteins 2006; 64:960–967.
65. Zhang Y. Template-based modeling and free modeling by I-TASSER in CASP7. Proteins 2007;69(Suppl 8):108–117.
66. Jauch R, Yeo H, Kolatkar PR, Clarke ND. Assessment of CASP7 structure predictions for template free targets. Proteins 2007;69 (Suppl 8):57–67.

67. Kopp J, Bordoli L, Battey JND, Kiefer F, Schwede T. Assessment of CASP7 predictions for template-based modeling targets. Proteins 2007;69(Suppl 8):38–56.
68. Read RJ, Chavali G. Assessment of CASP7 predictions in the high accuracy template-based modeling category. Proteins 2007;69 (Suppl 8):27–37.
69. Bordoli L, Kiefer F, Schwede T. Assessment of disorder predictions in CASP7. Proteins 2007;69(Suppl 8):129–136.
70. Tress M, Cheng J, Baldi P, Joo K, Lee J, Seo J-H, Lee J, Baker D, Chivian D, Kim D, Ezkurdia I. Assessment of predictions submitted for the CASP7 domain prediction category. Proteins 2007;69(Suppl 8):137–151.
71. Izarzugaza JMG, Graña O, Tress ML, Valencia A, Clarke ND. Assessment of intramolecular contact predictions for CASP7. Proteins 2007; 69(Suppl 8):152–158.
72. López G, Rajas A, Tress M, Valencia A. Assessment of predictions submitted for the CASP7 function prediction category. Proteins 2007;69(Suppl 8):165–174.
73. Greene LH, Lewis TE, Addou S, Cuff A, Dallman T, Dibley M, Redfern O, Pearl F, Nambudiry R, Reid A, Sillitoe I, Yeats C, Thornton JM, Orengo CA. The CATH domain structure database: new protocols and classification levels give a more comprehensive resource for exploring evolution. Nucleic Acids Res 2007;35 (Database Issue): D291–D297.
74. Clarke ND, Ezkurdia I, Kopp J, Read RJ, Schwede T, Tress M. Domain definition and target classification for CASP7. Proteins 2007; 69(Suppl 8):10–18.
75. Zemla A. LGA: a method for finding 3D similarities in protein structures. Nucleic Acids Res 2003;31:3370–3374.
76. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25:3389–3402.
77. Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, Church DM, DiCuccio M, Edgar R, Federhen S, Geer LY, Kapustin Y, Khovayko O, Landsman D, Lipman DJ, Madden TL, Maglott DR, Ostell J, Miller V, Pruitt KD, Schuler GD, Sequeira E, Sherry ST, Sirotkin K, Souvorov A, Starchenko G, Tatusov RL, Tatusova TA, Wagner L, Yaschenko E. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 2007;35 (Database Issue):D5–D12.
78. Martin AC, MacArthur MW, Thornton JM. Assessment of comparative modeling in CASP2. Proteins 1997;29(Suppl 1):14–28.
79. Peitsch MC. ProMod and Swiss-model: internet-based tools for automated comparative protein modelling. Biochem Soc Trans 1996;24:274–279.
80. Rychlewski L, Fischer D. LiveBench-8: the large-scale, continuous assessment of automated protein structure prediction. Protein Sci 2005;14:240–245.
81. Koh IY, Eyrich VA, Marti-Renom MA, Przybylski D, Madhusudhan MS, Eswar N, Grana O, Pazos F, Valencia A, Sali A, Rost B. EVA: evaluation of protein structure prediction servers. Nucleic Acids Res 2003;31:3311–3315.

WILEY InterScience® DISCOVER SOMETHING GREAT

PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

Template Based Assessment

Assessment of CASP7 predictions for template-based modeling targets

Jürgen Kopp,[1,2] Lorenza Bordoli,[1,2] James N.D. Battey,[1,2] Florian Kiefer,[1,2] and Torsten Schwede[1,2*]

[1] Biozentrum, University of Basel, Switzerland
[2] Swiss Institute of Bioinformatics, Basel, Switzerland

ABSTRACT

This manuscript presents the assessment of the template-based modeling category of the seventh Critical Assessment of Techniques for Protein Structure Prediction (CASP7). The accuracy of predicted protein models for 108 target domains was assessed based on a detailed comparison between the experimental and predicted structures. The assessment was performed using numerical measures for backbone and structural alignment accuracy, and by scoring correctly modeled hydrogen bond interactions in the predictions. Based on these criteria, our statistical analysis identified a number of groups whose predictions were on average significantly more accurate. Furthermore, the predictions for six target proteins were evaluated for the accuracy of their modeled cofactor binding sites. We also assessed the ability of predictors to improve over the best available single template structure, which showed that the best groups produced models closer to the target structure than the best single template for a significant number of targets. In addition, we assessed the accuracy of the error estimates (local confidence values) assigned to predictions on a per residue basis. Finally, we discuss some general conclusions about the state of the art of template-based modeling methods and their usefulness for practical applications.

Proteins 2007; 69(Suppl 8):38–56.
© 2007 Wiley-Liss, Inc.

***Key words:* CASP; template based protein modeling; comparative modeling; model quality assessment.**

INTRODUCTION

Protein structure modeling and prediction has gained significant interest in the biological research community for its ability to provide structural models for proteins lacking experimental structures. Template-based protein models, which exploit the evolutionary relationship between a target protein and others with known experimental structures, have been used successfully in a variety of applications, such as studying the effect of mutations, designing site-directed mutagenesis experiments, predicting binding sites, and docking small molecules in structure-based drug discovery. A variety of such modeling methods have been published over the last years. As the usefulness of a protein structure model depends on the accuracy of the prediction, it is crucial to identify the most suitable method for the task at hand from amongst this growing list of resources.

The Critical Assessment of Techniques for Protein Structure Prediction (CASP)[1] provides an objective evaluation of current prediction methods. In identifying their strengths and weaknesses this experiment serves two purposes. For biologists, this assessment aids in choosing the most suitable methods to meet their needs. For researchers engaged in the development of protein structure prediction techniques, detailed scrutiny of their methods and comparison with other approaches helps pinpoint strengths and limitations, and serves as a guide for future development. To ensure objectivity, CASP is organized as a double-blind prediction experiment, i.e. at

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

The authors state no conflict of interest.

*Correspondence to: Torsten Schwede, Klingelbergstrasse 50-70, 4056 Basel, Switzerland. E-mail: Torsten.Schwede@unibas.ch

Received 10 April 2007; Revised 23 July 2007; Accepted 26 July 2007
Published online 25 September 2007 in Wiley InterScience (www.interscience.wiley.com).
DOI: 10.1002/prot.21753

the time of the experiment, the predictors do not know the target structures, and the identity of the predictors is hidden from the assessors. At the end of the experiment, all predictions and assessment data are made publicly available. Accurate and appropriate assessment of protein structure prediction is not a simple standard procedure. While the main numerical assessment criteria have been well established over the series of CASPs,[1–4] progress and convergence in the prediction methods over subsequent CASP experiments require assessors to update existing criteria or even introduce new ones. This ensures that the assessment adequately appraises the overall quality of the models, as well as those features of the predictions that are relevant to their usefulness in specific scientific applications.

The assessment of the template-based modeling (TBM) category in CASP7 greatly benefits from having a broad basis for numerical and statistical analysis. With most participating groups having submitted predictions for the majority of target proteins, a sufficiently large number of diverse targets is available for comparing the various methods and testing the statistical significance of the differences between them. However, we would like to emphasize that this large number of targets, for which predictions had to be made within the 3 months of the prediction season in spring and summer 2006, represented an enormous work load for the participating predictor groups.

Predictions that could largely be built based on template structures were assessed in the TBM category. A subset of the TBM models was additionally assessed in the high accuracy (HA) category with respect to detailed structural features such as side-chain orientation and suitability for application in molecular replacement.[5] In this report, we present the results of our assessment of the models in the TBM category of CASP7 based on numerical criteria for evaluating the correctness of the overall structure and alignment, local residue interactions, accuracy of cofactor binding sites, and improvement over the best templates. Additionally, we evaluated the ability of predictors to correctly assign error estimates as per residue confidence values to their predictions. Finally, we discuss some general conclusions about the state of the art of template based modeling methods and their usefulness for practical applications.

RESULTS AND DISCUSSION

Targets, assessment units, and predictions in the TBM category

For the assessment, prediction target structures were split into assessment units (AU) and classified into three categories: free modeling (FM),[6] template-based modeling (TBM), and high-accuracy template-based models (HA).[5] Assessment units correspond to individual structural domains for single domain proteins. Multidomain proteins, for which the relative orientation could not be inferred from the template structure, were split into separate AUs. Multidomain proteins with the same relative orientation as the template were assessed as a single unit. Definition of assessment units and categorization criteria are discussed in detail elsewhere.[7] Traditionally, the term "target domain" has been used in previous CASP experiments to describe the segment of a prediction target to be assessed individually. However, depending on the context, "domain" denotes quite diverse concepts from an evolutionary, structural, or functional perspective. Therefore, we introduced the term "assessment unit" to describe the segments of a structure on which we based the assessment of 3D structure predictions. However, for historical reasons and for easier readability, the term "domain" might be used interchangeably with "assessment unit" in this manuscript.

In CASP7, the category of TBM comprises 108 out of a total of 123 assessment units, for which 15,717 predictions were submitted by 187 predictor groups. Sixty-eight groups were registered as prediction servers. Among the targets in the TBM category, 28 assessment units were also evaluated in the high-accuracy category (HA), and four overlapped with the definition of free modeling (FM) and were therefore assessed in both categories. The assessment units of the TBM category ranged in size from 526 residues for T0334, a flavin-dependent halogenase from *Nocardia aerocolonigenes*, to 36 residues for T0335, a NMR structure of protein ynzC from *Bacillus subtilis*. In Table I, we summarize the characteristics of the TBM targets. Predictions were received by the Protein Structure Prediction Center at UC Davis (in TS or AL format) and split according to assessment unit definitions. Standard numerical assessment data such as GDT-TS, GDT-HA, AL0, and RMSD values were provided to the assessors by the Protein Structure Prediction Center.

Assessment of the overall quality of the models

Visual inspection of all predictions showed that, like in previous CASP experiments, a significant number of physically impossible models were submitted. Models with more than 2% of the C_α atoms involved in clashes (C_α–C_α distance <1.9 Å) or more than 10% in bumps (C_α–C_α distance between 1.9 and 3.5 Å) and severely fragmented predictions were flagged as physically impossible. In total, 451 models were considered physically impossible. As shown in Figure 1, the majority of these models were submitted by only a few groups, while the distribution over assessment units is homogenous. Since three target structures (T0320, T0332, T0378) were methyltransferases containing topological knots,[8] predictions with knots were not penalized, provided that the structures passed the C_α–C_α distance criteria outlined before.

Numerical assessment in recent CASPs has been based on the two well-established criteria, GDT and AL0. AL0 is defined as the percentage of correctly aligned residues

Table I
TBM Assessment Units in CASP7

Target	Residues	UniProt	Description	Best template	LGA-S	Seq. id
T0283	97	Q9K5V7	JCSG target 10176605, BH3980 protein, *Bacillus halodurans*	2b2jA	54.0	6.7
T0284	250	Q9HUU1	Member of isocitrate lyase family, *Pseudomonas aeruginosa*	1oqfA	87.8	30.1
T0285	99	n/a	Extracytoplasmic domain from histidine kinase, *Cellvibrio japonicus*	1p0zG	54.6	10.1
T0286	202	A3DDK4	Lipolytic enzyme, *Clostridium thermocellum*	1esd	76.7	20.1
T0288	86	Q9NRD5	PDZ domain of PICK1, *Homo sapiens*	2fneB	93.4	32.5
T0289_1	233	Q9R1T5	Aspartoacylase, *Rattus norvegicus*	1yw4A	55.8	19.4
T0289_2	74			1vdzA	59.5	9.6
T0290	173	Q13427	Peptidyl-prolyl isomerase domain of cyclophilin G, *Homo sapiens*	1c5fM	99.2	60.5
T0291	281	P29320	Epha3 Receptor Tyrosine Kinase and Juxtamembrane Region, *Homo sapiens*	1jpaA	92.8	81.1
T0292_1	77	P51955	Nek2 Centrosomal Kinase, *Homo sapiens*	2bmcF	95.4	33.8
T0292_2	173			2acxA	85.7	28.5
T0293	198	Q86W50	Methyltransferase 10 domain-containing protein, *Homo sapiens*	1nv9A	66.2	20.9
T0295_1	180	Q8ILT8	Dimethyladenosine transferase, *Plasmodium falciparum*	1zq9B	95.2	49.2
T0295_2	95			1zq9A	96.9	41.1
T0297	211	Q97PY9	Putative platelet activating factor, *Streptococcus pneumoniae*	1bwp	72.8	22.6
T0298_1	148	O87014	Putative aspartate-semialdehyde dehydrogenase, *Pseudomonas aeruginosa*	2g17A	82.7	21.2
T0298_2	186			1pquA	88.9	16.6
T0299_1	91	Q97RI5	MCSG target APC80351, *Streptococcus pneumoniae*	2cg8C	68.6	13.7
T0299_2	89			1rjjA	60.2	9.1
T0301_1	200	Q9I5E5	JCSG target np_249484.1, *Pseudomonas aeruginosa*	1w61A	53.4	12.3
T0301_2	191			1w62A	47.0	11.8
T0302	129	Q9NS28	RGS domain of RGS18, *Homo sapiens*	1agrE	90.5	53.9
T0303_1	147	Q0I1W8	Phosphoglycolate phosphatase, *Haemophilus somnus*	2ah5A	89.0	28.4
T0303_2	77			1fezB	81.3	7.1
T0304	101	P76364	YeeU protein, *Escherichia coli*	2gnxA	55.9	8.3
T0305	280	P23470	Tyrosine receptor phosphatase gamma, *Homo sapiens*	2fh7A	95.9	53.1
T0306	95	P0AEJ8	Ethanolamine utilization protein, *Escherichia coli*	1d7qA	55.6	15.8
T0308	165	Q9H0F7	ADP-ribosylation factor-like protein 6, *Homo sapiens*	1o3yB	93.7	41.5
T0311	64	P67699	Antitoxin HigA, *Escherichia coli*	1rpeL	90.0	18.0
T0312	132	O30132	NESG target GR103, Archaeoglobus fulgidus	1xv2B	61.1	13.5
T0313	316	Q9BVG8	Kinesin-like protein KIFC3 motor domain, *Homo sapiens*	1ii6A	88.4	42.1
T0315	253	Q7A1S8	TatD deoxyribonuclease, *Staphylococcus aureus*	1j6oA	94.9	39.0
T0316_1	188	Q97T38	tRNA (5-Methylaminomethyl-2-Thiouridylate)-Methyltransferase TrmU,	1kh3C	51.0	17.4
T0316_3	90		*Streptococcus pneumoniae*	1wb3B	78.1	17.9
T0317	149	Q8BTR5	putative dual specificity phosphatase, *Mus musculus*	2esbA	92.4	34.5
T0318_1	154	P34629	Leucine aminopeptidase, *Caenorhabditis elegans*	1vhuA	38.9	4.4
T0318_2	335			1gytL	88.7	28.7
T0320_1	214	P38913	FAD synthetase, *Saccharomyces cerevisiae*	1sur	55.0	19.3
T0321_1	96	Q18YZ7	JCSG target ZP_00559375.1, *Desulfitobacterium hafniense*	1f9cA	59.3	18.8
T0321_2	148			1kxzE	48.0	8.8
T0322	128	P25734	Colonization factor antigen I subunit E, *Caulobacter crescentus*	2h4uD	85.4	20.8
T0323_1	101	Q9KC25	DNA-3-methyladenine glycosidase, *Bacillus halodurans*	1yqmA	55.7	20.8
T0323_2	116			1dizA	87.8	26.6
T0324_1	142	Q88YA8	Putative phosphoglycolate phosphatase, *Lactobacillus plantarum*	2fdrA	89.6	25.0
T0324_2	65			2ah5A	92.6	4.7
T0325	261	P59745	Protein EF3048, *Enterococcus faecalis*	1v6tA	49.1	16.1
T0326	289	Q9WZY3	Homoserine O-succinyltransferase, *Thermotoga maritima*	2ghrA	84.9	55.2
T0327	73	O31639	YjcQ protein, *Bacillus subtlis*	1lnwF	76.5	13.3
T0328	307	Q8EIU4	Putative melanin biosynthesis protein TyrA, *Shewanella oneidensis*	2gvkA	90.5	30.4
T0329_1	141	Q1GA24	Putative phosphoglycolate phosphatase, *Lactobacillus delbrueckii*	1rdfB	90.9	25.8
T0329_2	92			1rqlA	60.7	11.6
T0330_1	153	Q8KBS5	Haloacid dehalogenase-like hydrolase, *Chlorobium tepidum*	2ah5A	82.9	29.2
T0330_2	72			1lvhB	73.5	7.9
T0331	139	Q2ZZ07	Pyridoxamine 5'-phosphate oxidase-related protein, *Streptococcus suis*	1ty9A	72.3	16.4
T0332	153	Q13395	Methyltransferase Domain of Human TAR (HIV-1) RNA binding protein 1, *Homo sapiens*	1zjrA	88.8	22.9
T0333_1	206	Q8KND7	CalG3, *Micromonospora echinospora*	1rvvB	48.9	15.7
T0333_2	148			1rvvB	69.2	25.6
T0334	526	Q8KHZ8	Flavin-dependent halogenase, *Nocardia aerocolonigenes*	2ajqA	94.4	56.2
T0335	36	O31818	YnzC protein, *Bacillus subtilis*	1yluA	96.7	0.0
T0338_1	143	O60583	Cyclin T2, *Homo sapiens*	1jkw	74.6	20.3
T0338_2	113			1n4mA	60.9	10.0

(*Continued*)

Table I
Continued

Target	Residues	UniProt	Description	Best template	LGA-S	Seq. id
T0339_1	136	Q7L670	Selenocysteine lyase, *Homo sapiens*	1eg5B	78.3	26.1
T0339_2	267			1eg5A	87.3	37.3
T0340	82	Q15599	Second PDZ domain of human NHERF-2, *Homo sapiens*	1g9oA	98.0	59.8
T0341_1	148	Q6PEB2	Haloacid dehalogenase-like hydrolase domain containing protein, *Mus musculus*	1zjjB	90.0	27.3
T0341_2	104			1wviB	93.8	21.4
T0342	122	O75223	Protein LOC79017, *Homo sapiens*	2g0qA	80.4	21.5
T0345	185	P18283	Glutathionine peroxidase 2, *Homo sapiens*	1gp1A	97.0	68.1
T0346	172	P30414	Peptidylprolyl isomerase domain of the human NK-tumour recognition protein, *Homo sapiens*	2gw2A	99.9	71.5
T0347_1	89	Q8UF59	Protein Atu1540, Agrobacterium tumefaciens	1vk1A	74.3	21.4
T0348	61	Q7NSS5	Putative Tetraacyldisaccharide-1-P 4-kinase, *Chromobacterium violaceum*	1rfs	58.1	16.2
T0349	57	Q6NAY9	Protein RPA1041, *Pseudomonas aeruginosa*	1yj7D	87.4	17.3
T0351	56	P54342	Phage-like element PBSX protein xkdW, *Bacillus subtilis*	1cs1C	62.8	4.6
T0354	122	Q7P0P8	Protein CV0518, *Chromobacterium violaceum*	2be3A	56.8	8.1
T0356_2	192	P0AAB4	3-octaprenyl-4-hydroxybenzoate decarboxylase, *Escherichia coli*	1ejeA	45.6	9.6
T0357	132	O30181	NESG Target GR101, *Archaeoglobus fulgidus*	1aco	56.3	13.5
T0358	65	P75677	Protein ykfF, *Escherichia coli*	1dgd	60.5	13.0
T0359	90	O75970	3rd PDZ domain of multiple pdz domain protein MPDZ, *Homo sapiens*	2bygA	92.9	31.0
T0360	97	Q9JY98	Protein NMB1681, *Neisseria meningitidis*	1dvoA	76.6	21.0
T0362	144	Q7V4A7	JCSG target NP_895880.1, *Prochlorococcus marinus*	2gf6A	82.9	22.7
T0363	46	Q4QNE7	NYSGRC 68057197, *Haemophilus influenzae*	2bb6A	79.9	13.2
T0364	147	Q88R33	JCSG target NP_742468.1, *Pseudomonas putida*	2av9B	79.7	14.3
T0365	207	Q8EAX1	JCSG target NP_719307.1, *Shewanella oneidensis*	1xwmA	67.0	13.1
T0366	84	O75970	12th PDZ domain of multiple pdz domain protein MPDZ, *Homo sapiens*	2fneB	93.5	26.8
T0367	123	O29944	JCSG target NP_069135.1, *Archaeoglobus fulgidus*	1ufbC	87.1	15.0
T0368	157	Q8KAL8	JCSG target NP_663012.1, *Chlorobium tepidum*	2c2lC	69.4	17.0
T0369	147	Q41IB9	JCSG target ZP_00537729.1, *Exiguobacterium sibiricum*	1rxqA	61.8	11.4
T0370	144	Q825J7	JCSG target NP_828636.1, *Streptomyces avermitilis*	1vl7B	75.4	18.8
T0371_1	162	Q11S56	Puatative HAD superfamily sugar phosphatase, *Cytophaga hutchinsonii*	1vjrA	80.7	27.7
T0371_2	121			1zjjA	65.6	20.2
T0372_1	126	Q8A1H2	Protein BT_3689, *Bacteroides thetaiotaomicron*	1ro5A	59.2	4.0
T0372_2	172			1xf8A	60.1	9.1
T0373	140	Q9HZE1	Putative transcriptional regulator protein, *Pseudomonas aeruginosa*	1s3jB	70.6	23.8
T0374	160	Q9HV14	Putative acetyltransferase, *Pseudomonas aeruginosa*	1tiqA	75.2	9.4
T0375	296	P50053	Ketohexokinase, *Homo sapiens*	2fv7B	67.5	18.9
T0376	306	Q8U6Y1	Dihydrodipicolinate synthase, *Agrobacterium tumefaciens*	1xxxB	81.3	25.0
T0378_1	89	Q7MW92	Putative RNA methyltransferase of the TrmH family, *Porphyromonas gingivalis*	1ipaA	74.1	22.0
T0378_2	142			1gz0C	89.6	27.4
T0379_1	140	Q7MWA6	Putative HAD-like family hydrolase, *Porphyromonas gingivalis*	1zd5A	83.9	27.8
T0379_2	64			2b0cA	67.2	25.5
T0380	142	Q97DI6	Pyridoxinephosphate oxidase family-related protein, *Clostridium acetobutylicum*	2fhqA	82.7	25.8
T0381_1	61	Q0SH23	Putative transcriptional regulator RHA06195, *Rhodococcus sp. RHA1*	2g7uA	99.5	45.9
T0381_2	176			2g7uD	92.7	39.8
T0382	119	Q6N8L4	MCSG target APC6185, *Rhodopseudomonas palustris*	1kpsB	56.8	9.8
T0383	125	Q97PP5	Protein SP_1558, *Streptococcus pneumoniae*	1qynB	63.2	11.8
T0384	301	Q97PV8	Gfo/Idh/MocA family Oxidoreductase, *Streptococcus pneumoniae*	1ydwA	84.2	22.2
T0385	125	O05815	Protein rv2844, *Mycobacterium tuberculosis*	1jgcB	87.2	10.9
T0386_1	206	Q6G2A9	Putative cell filamentation protein, *Bartonella henselae*	2g03A	63.8	24.5

in the 5 Å LGA sequence-independent superposition of the model and experimental structure of the target. A model residue is considered to be correctly aligned if the predicted C_α atom position falls within 3.8 Å of the corresponding experimental atom, and there is no other C_α atom of the experimental structure nearer. GDT (global distance test) identifies sets of residues in the predictions deviating from the target by not more than a specified C_α distance cutoff for different sequence-dependent superpositions, e.g. using distance cut-off values of 1, 2, 4, and 8 Å for GDT-TS calculation. It was suggested during the CASP6 meeting in Gaeta that the cut-off values applied to calculate GDT-TS may not be appropriate to detect small differences in backbone quality.[2] In our assessment, we considered the upper cut-off value of 8 Å as too lenient to discriminate the finer structural differences between models of template-based predictions and therefore decided to use GDT-HA with distance cut-off values of 0.5, 1, 2, and 4 Å in our evaluation.

Although GDT is a sequence-dependent and AL0 a sequence-independent measure, both scores are highly correlated and on average contain little complementary

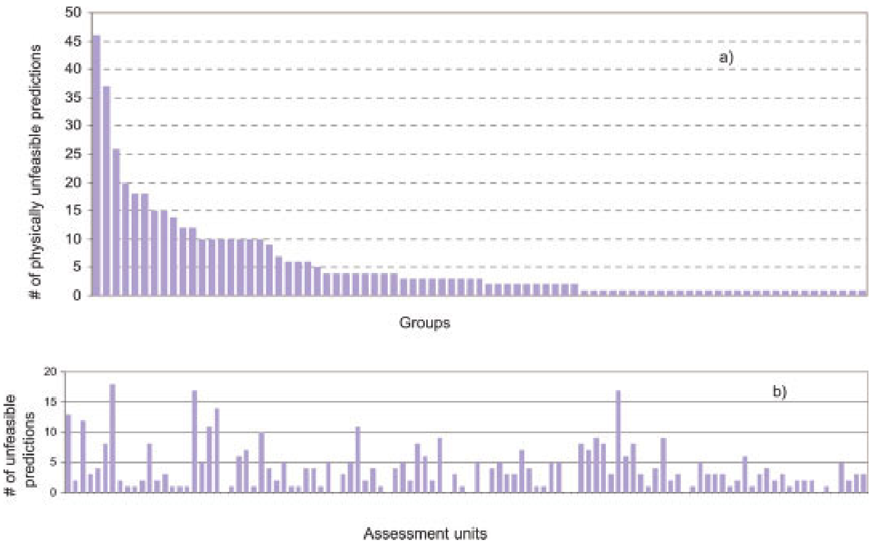


Figure 1

Distribution of physically impossible models. The majority of unfeasible models were submitted by only a few groups (a), while the distribution over assessment units is homogenous (b). [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

information for comparing and contrasting the different prediction methods (Fig. 2). As both GDT and AL0 are derived from global superpositions of C_α coordinates only, they do not reflect important local structural features of a protein such as backbone geometry, packing of amino acid side-chains, and atomic interactions like hydrogen bonds or hydrophobic contacts. To complement these global C_α-based criteria, we introduced a local atomic measure termed HBscore. It counts the intersection of corresponding hydrogen bonds present in the model and the target structure: HBscore = number of (H-bonds in model $\cap$ H-bonds in target structure)/number of H-bonds in target structure. We excluded hydrogen bonds involving side-chain atoms of residues with more than 50% relative surface exposure in the target structure. In the predicted structures, hydrogen bonds were not considered if they involved amino acid residues with incorrect topology or had severe clashes with neighboring residues ($d < 1.2$ Å). Hydrogen bonds were calculated using HBPlus,[9] and relative solvent accessibility of side-chains using NACCESS (Hubbard and Thornton, 1993). The HBscore enumeration of specific H-bond interactions accounts for ambiguities arising from chemically equivalent side-chain atoms being assigned different atom names in IUPAC nomenclature (e.g., Glu OE1, OE2; Arg NH1, NH2, etc.). Figure 3 illustrates HBscore for the example of a short β-sheet. When comparing structure predictions and actual experimental structures, high scores in global criteria such as GDT and AL0 are necessary, but not by themselves sufficient indicators of accuracy. Local criteria based on specific atomic interactions provide complementary information, as illustrated in Figure 4.

Numerical evaluation and statistical significance of the results

The assessment of the individual groups was based on the predictions submitted as "Model 1." The majority of groups predicted more than one hundred assessment units and consequently, for each target more than 100

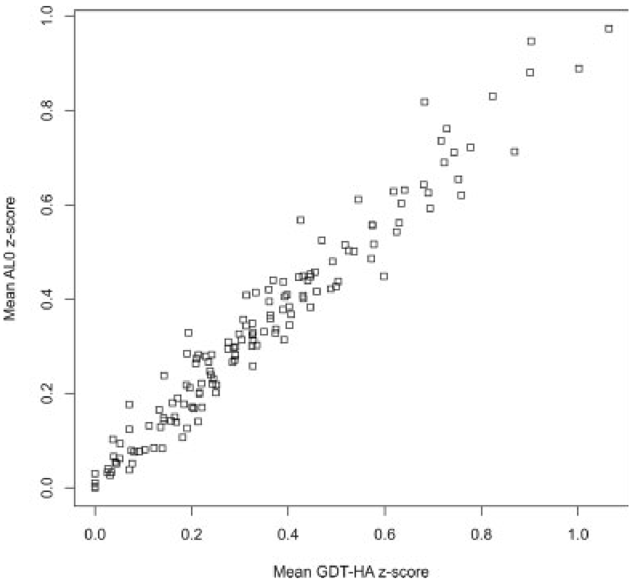
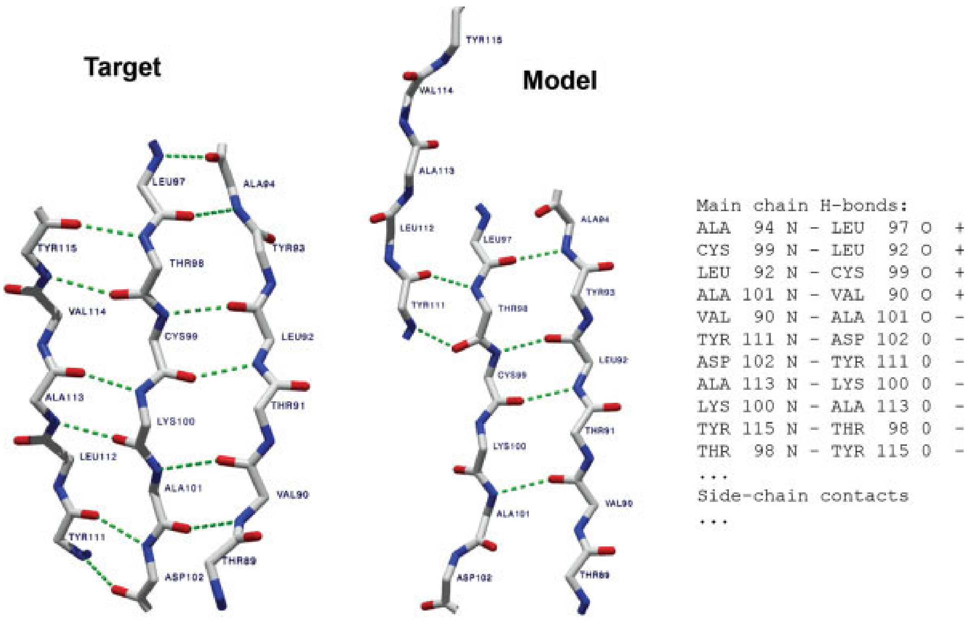


Figure 2

Correlation of GDT-HA and AL0 z-scores for groups in CASP7.



```
Main chain H-bonds:
ALA  94 N - LEU  97 O  +
CYS  99 N - LEU  92 O  +
LEU  92 N - CYS  99 O  +
ALA 101 N - VAL  90 O  +
VAL  90 N - ALA 101 O  -
TYR 111 N - ASP 102 0  -
ASP 102 N - TYR 111 0  -
ALA 113 N - LYS 100 0  -
LYS 100 N - ALA 113 0  -
TYR 115 N - THR  98 0  -
THR  98 N - TYR 115 0  -
...
Side-chain contacts
...
```

Figure 3

Illustration of HBscore on the example of a short β-sheet structure.

individual predictions were available for the assessment. For groups which submitted both unrefined and refined models, the assessment was based on the refined prediction. If predictions for a target were submitted in several fragments, the segment with the longest overlap with the assessment unit was assessed.

The scoring scheme adopted in our assessment was similar to the one used in previous CASP experiments.[2,3] To allow the comparison of the results for targets of different modeling difficulty, we computed *z*-scores for both GDT-HA and AL0 for each assessment unit in the following way: (i) For each assessment unit, average values and standard deviations for all predictions were calculated; (ii) For those predictions, whose scores were not more than two standard deviations below average and which were not flagged as physically impossible, we recomputed the means and standard deviations, and used these to assign *z*-scores to all predictions; (iii) Models

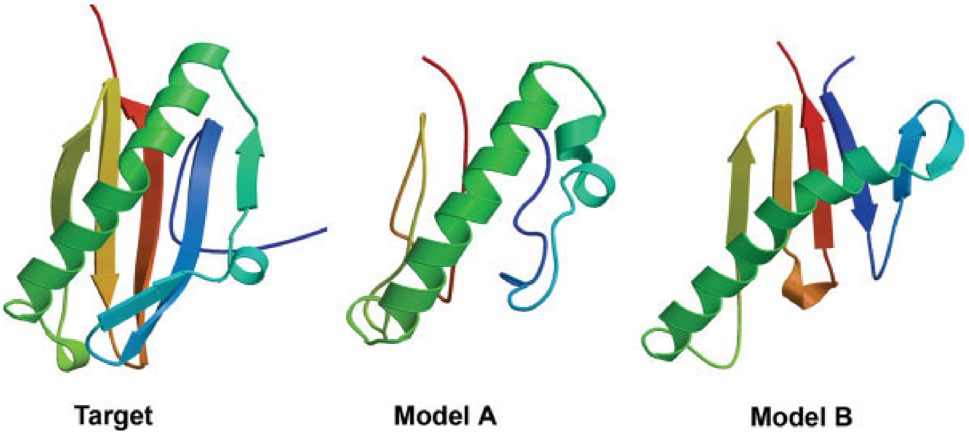


Figure 4

Complementary information assessed by GDT-TS and HBscore. While model A better resembles the global positioning of Cα atoms of the target structure (GDT-TS: 46) compared to model B (GDT-TS: 35), model B better reflects the H-bonding pattern in the central β-sheet structure (HBscore: 41) compared to model A (HBscore: 26).

Table II
Mean Values and z-Scores for Individual Prediction Groups

Group		Group name	Number of predictions	Mean GDT-HA	Mean AL0	Mean GDT-HA *z*-score	Mean AL0 *z*-score	Combined GDT-HA and AL0 *z*-score
4	s	ROBETTA	107	46.77	60.14	0.54	0.50	0.52
5		luethy	108	47.83	65.46	0.68	0.82	0.75
9		CBiS	4	5.72	0.00	0.00	0.00	n.a.
10		SAM-T06	108	44.78	56.85	0.40	0.41	0.40
11		Dlakic-MSU	10	53.24	68.94	0.26	0.17	n.a.
13		Jones-UCL	107	47.19	62.24	0.52	0.50	0.51
15		Advanced-ONIZUKA	35	20.68	10.43	0.22	0.17	0.20
16		AMBER/PB	1	54.26	88.37	0.00	0.15	n.a.
18		LUO	97	45.31	58.99	0.52	0.52	0.52
20		Baker	106	50.19	65.21	1.00	0.89	0.95
21		karypis	83	39.02	49.23	0.37	0.34	0.36
22	s	karypis.srv	106	38.28	48.87	0.23	0.27	0.25
24		Zhang	107	51.49	67.87	1.06	0.97	1.02
25	s	Zhang-Server	108	50.35	66.60	0.90	0.88	0.89
26		SAMUDRALA	106	47.87	61.55	0.69	0.63	0.66
27		SAMUDRALA-AB	106	46.81	60.09	0.57	0.56	0.57
28	s	PROTINFO	103	44.34	55.99	0.33	0.31	0.32
29	s	PROTINFO-AB	106	42.61	53.91	0.31	0.36	0.33
30		TsaiLab	52	45.86	60.48	0.22	0.22	0.22
31		Avbelj	7	15.72	0.32	0.00	0.00	n.a.
33		POEM-REFINE	19	28.66	25.75	0.47	0.43	n.a.
34		ROKKO	105	43.73	56.30	0.40	0.38	0.39
35	s	ROKKY	106	42.94	53.56	0.28	0.31	0.29
38		GeneSilico	102	48.34	63.61	0.72	0.74	0.73
40		YASARA	23	56.05	74.35	0.36	0.42	0.39
43		hu	2	54.75	69.09	0.00	0.00	n.a.
44	s	gtg	56	38.85	49.75	0.10	0.08	0.09
45		INFSRUCT	1	10.17	0.00	0.00	0.00	n.a.
46	s	Pcons6	108	46.70	60.70	0.50	0.44	0.47
47	s	Pmodeller6	108	46.98	61.06	0.58	0.52	0.55
50		SBC	105	49.05	65.18	0.78	0.72	0.75
54		PROTEO	62	8.00	1.29	0.00	0.00	0.00
60		HIT-ITNLP	104	27.98	32.74	0.04	0.05	0.05
62		Floudas	21	26.24	19.08	0.19	0.13	0.16
63		FEIG	106	36.80	48.75	0.14	0.24	0.19
64		LMM-Bicocca	29	42.37	50.29	0.21	0.14	0.18
65		Protofold	2	9.64	0.00	0.00	0.00	n.a.
66		UF_GATORS	4	16.74	18.15	0.00	0.00	n.a.
69	s	panther2	78	33.10	41.68	0.05	0.06	0.06
71		Wymore	45	40.34	51.48	0.13	0.17	0.15
74		SHORTLE	91	45.97	58.28	0.36	0.37	0.36
78		Dill-ZAP	5	26.52	10.12	0.01	0.03	n.a.
83	s	LOOPP	108	41.89	51.37	0.25	0.23	0.24
87		Pan	108	42.91	52.91	0.30	0.31	0.31
91		Ma-OPUS	107	44.82	56.86	0.37	0.33	0.35
92	s	Ma-OPUS-server	108	43.17	53.69	0.35	0.33	0.34
102	s	Huber-Torda-Server	102	39.58	47.65	0.20	0.17	0.19
103		Huber-Torda	107	42.46	52.14	0.25	0.20	0.23
105		andante	106	47.28	60.12	0.63	0.56	0.60
109		Cracow.pl	40	12.15	1.03	0.00	0.00	0.00
111		panther	82	49.27	67.44	0.21	0.26	0.24
113		Bates	108	47.41	62.36	0.63	0.60	0.62
121		Peter-G-Wolynes	24	17.35	9.26	0.07	0.18	0.12
125		TASSER	108	49.89	66.68	0.90	0.95	0.92
132		Softberry	102	39.60	51.29	0.19	0.28	0.24
135		CBSU	108	43.34	54.48	0.33	0.35	0.34
136	s	FOLDpro	108	44.97	56.46	0.45	0.38	0.41
137	s	3Dpro	107	45.42	57.02	0.49	0.42	0.45
139	s	ABIpro	107	14.38	6.56	0.03	0.03	0.03
168	s	Distill	108	26.34	30.57	0.04	0.07	0.05
170		LMU	78	44.99	55.92	0.12	0.08	0.10
174		Bystroff	58	28.71	30.35	0.08	0.08	0.08

(*Continued*)

Table II
Continued

Group		Group name	Number of predictions	Mean GDT-HA	Mean AL0	Mean GDT-HA *z*-score	Mean AL0 *z*-score	Combined GDT-HA and AL0 *z*-score
178		Bilab	108	42.85	53.83	0.33	0.26	0.29
179	s	Bilab-ENABLE	107	41.43	52.19	0.25	0.22	0.23
186	s	CaspIta-FOX	107	41.32	51.76	0.22	0.20	0.21
191		Schomburg-group	22	56.25	75.99	0.57	0.49	0.53
193	s	karypis.srv.4	91	9.47	1.38	0.00	0.00	0.00
194		Scheraga	34	16.15	4.80	0.03	0.04	0.03
197		MTUNIC	103	27.91	33.33	0.07	0.12	0.10
203		forecast	103	34.68	41.59	0.16	0.14	0.15
205		NanoModel	108	40.31	51.31	0.21	0.27	0.24
208		Nano3D	63	42.51	53.89	0.29	0.30	0.29
209		NanoDesign	89	46.10	58.70	0.33	0.30	0.32
211		KIST	103	41.55	53.03	0.23	0.28	0.25
212	s	HHpred1	108	47.00	61.55	0.57	0.56	0.57
213	s	HHpred2	108	48.43	62.25	0.76	0.62	0.69
214	s	BayesHH	108	47.40	61.21	0.62	0.54	0.58
224		ricardo	4	38.99	54.45	0.66	0.72	n.a.
226		Struct-Pred-Course	2	43.86	60.91	0.10	0.00	n.a.
234		McCormack_Okazaki	10	40.74	48.92	0.23	0.21	n.a.
239	s	nFOLD	108	41.66	51.07	0.22	0.20	0.21
242	s	FUGUE	105	42.46	53.02	0.20	0.17	0.19
243	s	UNI-EID_sfst	104	46.33	61.00	0.39	0.40	0.40
245	s	UNI-EID_expm	107	47.17	61.82	0.33	0.32	0.32
247	s	3D-JIGSAW_POPULUS	104	39.50	48.07	0.14	0.15	0.15
248	s	RAPTOR	108	45.09	58.04	0.43	0.45	0.44
249		taylor	39	27.66	27.81	0.17	0.14	0.15
250		fleil	77	39.91	52.24	0.14	0.13	0.13
252		EAtorP	15	12.66	2.33	0.00	0.00	n.a.
257	s	FORTE1	108	37.65	46.62	0.14	0.14	0.14
261	s	mGen-3D	107	43.66	55.49	0.33	0.30	0.31
263		igor	45	13.60	3.54	0.00	0.03	0.02
267	s	RAPTOR-ACE	108	45.93	60.01	0.46	0.46	0.46
268	s	karypis.srv.2	108	37.12	46.51	0.19	0.22	0.20
273		BioDec	70	35.80	47.84	0.04	0.10	0.07
274	s	shub	107	44.95	58.57	0.33	0.41	0.37
275	s	beautshot	108	45.55	59.45	0.39	0.44	0.41
276		keasar	107	44.08	59.57	0.37	0.44	0.40
277	s	keasar-server	101	41.79	55.30	0.19	0.33	0.26
278		Pushchino	4	16.43	14.92	0.02	0.03	n.a.
284		Oka	4	23.78	28.93	0.02	0.03	n.a.
297		MLee	101	42.50	52.95	0.31	0.34	0.33
298	s	CIRCLE	108	46.60	61.17	0.47	0.52	0.50
302	s	3D-JIGSAW	104	38.36	46.66	0.08	0.08	0.08
304	s	panther3	16	33.54	42.68	0.03	0.08	n.a.
307	s	MetaTasser	108	45.44	60.61	0.62	0.63	0.62
316	s	FORTE2	108	37.17	45.94	0.17	0.15	0.16
318	s	FUNCTION	107	45.01	57.68	0.33	0.33	0.33
319	s	FUGMOD	100	42.99	54.14	0.24	0.22	0.23
333	s	forecast-s	98	36.48	45.83	0.16	0.18	0.17
337		AMU-Biology	98	45.76	57.80	0.41	0.37	0.39
338		UCB-SHI	103	47.20	59.59	0.50	0.43	0.46
347	s	beautshotbase	106	46.30	58.70	0.40	0.34	0.37
349	s	FAMSD	108	46.27	59.31	0.45	0.45	0.45
351	s	FAMS	108	46.21	60.03	0.44	0.44	0.44
361		Doshisha-Nagoya	7	17.90	3.82	0.00	0.00	n.a.
368	s	Frankenstein	56	35.92	42.73	0.14	0.08	0.11
380	s	SAM-T99	87	47.82	63.20	0.28	0.27	0.28
381	s	SAM-T02	104	43.17	55.00	0.24	0.24	0.24
383	s	UNI-EID_bnmx	108	46.29	60.22	0.49	0.48	0.49
389	s	SAM_T06_server	108	42.28	52.38	0.29	0.27	0.28
393		Distill_human	108	26.41	30.35	0.04	0.05	0.05
397		Tripos-Cambridge	10	58.42	74.17	0.31	0.23	n.a.
401		MIG	90	39.88	46.71	0.18	0.11	0.14
413	s	SPARKS2	108	45.10	57.62	0.39	0.38	0.38

(*Continued*)

Table II
Continued

Group		Group name	Number of predictions	Mean GDT-HA	Mean AL0	Mean GDT-HA *z*-score	Mean AL0 *z*-score	Combined GDT-HA and AL0 *z*-score
414	s	SP3	108	46.07	59.13	0.46	0.42	0.44
415	s	SP4	108	45.44	58.31	0.43	0.40	0.42
416		honiglab	100	46.57	60.06	0.42	0.45	0.43
418	s	HHpred3	108	47.81	61.45	0.69	0.59	0.64
420	s	3D-JIGSAW_RECOM	104	39.32	47.35	0.09	0.08	0.08
427		CADCMLAB	102	21.78	18.45	0.03	0.03	0.03
435	s	RAPTORESS	108	43.70	57.44	0.31	0.41	0.36
437		osgdj	11	15.81	5.53	0.00	0.00	n.a.
439		Sternberg	107	45.85	60.11	0.45	0.45	0.45
443		fais	79	35.67	43.46	0.21	0.28	0.25
453		Deane	18	17.99	14.93	0.09	0.09	n.a.
464	s	POMYSL	52	11.69	1.70	0.00	0.01	0.01
468	s	Phyre-1	104	41.08	52.90	0.20	0.21	0.20
469	s	Phyre-2	107	43.19	55.77	0.30	0.33	0.31
474		PUT_lab	72	28.73	28.74	0.08	0.05	0.06
483		Hirst-Nottingham	13	15.35	2.20	0.00	0.00	n.a.
490		lwyrwicz	106	43.49	55.20	0.29	0.30	0.30
494	s	CPHmodels	60	47.06	61.96	0.11	0.13	0.12
495		largo	2	52.36	76.88	1.02	1.08	n.a.
501		Bristol_Comp_Bio	4	60.84	79.70	0.06	0.06	n.a.
509		SEZERMAN	66	30.18	32.69	0.03	0.03	0.03
511	s	FPSOLVER-SERVER	103	9.67	0.81	0.00	0.00	0.00
527		chaos	16	37.19	47.85	0.07	0.21	n.a.
536		Chen-Tan-Kihara	103	44.08	55.72	0.36	0.36	0.36
550		ZIB-THESEUS	94	29.90	30.93	0.05	0.09	0.07
556		LEE	106	49.16	62.17	0.87	0.71	0.79
559		GSK-CCMM	4	66.42	87.97	0.45	0.49	n.a.
564		ShakSkol-AbInitio	13	28.86	29.12	0.53	0.57	n.a.
568		CHIMERA	107	49.21	64.85	0.73	0.76	0.74
586	s	MIG_FROST	47	29.20	28.49	0.07	0.04	0.05
588	s	MIG_FROST_FLEX	2	38.56	44.13	0.32	0.08	n.a.
599		KORO	31	21.48	15.14	0.43	0.57	0.50
601		LTB-WARSAW	86	43.29	55.09	0.28	0.29	0.28
609	s	GeneSilicoMetaServer	100	46.57	59.99	0.43	0.41	0.42
610		Dlakic-DGSA	3	51.11	73.71	0.00	0.22	n.a.
614		Brooks_caspr	21	48.72	63.11	0.60	0.45	0.52
638		Soeding	1	29.93	36.36	1.36	1.29	n.a.
640		jive	99	40.61	51.98	0.24	0.28	0.26
641		tlbgroup	14	51.54	68.97	0.18	0.24	n.a.
650		Schulten	15	43.25	52.10	0.43	0.36	n.a.
651		verify	108	48.35	63.41	0.68	0.64	0.66
654	s	NN_PUT_lab	103	43.00	53.86	0.24	0.25	0.24
658		hPredGrp	107	49.15	64.19	0.72	0.69	0.71
659		CHEN-WENDY	32	63.93	81.46	0.39	0.31	0.35
664		CIRCLE-FAMS	108	48.95	64.51	0.74	0.71	0.73
671		fams-multi	108	48.51	62.92	0.64	0.63	0.64
673		ProteinShop	6	18.32	2.74	0.00	0.18	n.a.
675		fams-ace	108	49.64	65.79	0.82	0.83	0.83
677		UAM ICO BIB	96	41.95	52.80	0.36	0.39	0.38
683		MUMSSP	15	64.17	82.80	0.34	0.38	n.a.
698		MQAP-Consensus	108	49.04	64.01	0.75	0.65	0.70
705		Akagi	101	36.96	46.28	0.17	0.19	0.18
706		TENETA	106	39.39	49.01	0.18	0.18	0.18
710		Ligand-Circle	94	46.13	59.29	0.54	0.61	0.58
721		ROBETTA-late	3	33.48	44.76	0.35	0.40	n.a.
728	s	Ma-OPUS-server2	71	42.47	52.02	0.29	0.28	0.29
735		EBGM	13	29.60	36.03	0.01	0.05	n.a.
736		dokhlab	18	31.10	29.24	0.13	0.16	n.a.
746		CDAC	4	13.25	0.82	0.00	0.00	n.a.
757		SSU	16	22.33	13.34	0.11	0.17	n.a.
781		SCFBio-IITD	2	27.14	0.00	0.00	0.00	n.a.
794		MerzShak	4	27.22	27.86	0.70	0.70	n.a.

n.a.: Groups with less than 20 predictions were not included in the final ranking.

Table III

Statistical Significance of the Results of the 25 Highest Scoring Groups

	24	20	125	25	675	556	5	50	568	664	38	658	698	213	651	26	418	671	307	113	105	214	710	212	27
24		G: 0.74(0.24) A: 1.73(0.12) H:-6.87(0.00)	G: 1.46(0.00) A: 1.23(0.15) H: 3.79(0.00)	G: 1.10(0.00) A: 1.34(0.00) H:-0.38(0.50)	G: 1.73(0.00) A: 1.69(0.02) H: 2.77(0.00)	G: 1.62(0.00) A: 4.08(0.00) H:-3.50(0.00)	G: 2.75(0.00) A: 1.77(0.01) H: 2.06(0.00)	G: 2.25(0.00) A: 3.24(0.00) H:-2.15(0.00)	G: 2.09(0.00) A: 2.56(0.00) H: 0.58(0.36)	G: 2.31(0.00) A: 3.13(0.00) H:-1.71(0.02)	G: 2.71(0.00) A: 3.45(0.00) H:-3.34(0.00)	G: 2.38(0.00) A: 3.37(0.00) H: 1.00(0.16)	G: 2.15(0.00) A: 3.55(0.00) H:-0.30(0.63)	G: 2.33(0.00) A: 4.38(0.00) H:-0.36(0.67)	G: 2.73(0.00) A: 3.75(0.00) H: 0.23(0.73)	G: 2.73(0.00) A: 4.69(0.00) H:-1.80(0.02)	G: 2.81(0.00) A: 4.83(0.00) H: 0.20(0.83)	G: 2.88(0.00) A: 4.44(0.00) H: 2.34(0.00)	G: 3.30(0.00) A: 4.36(0.00) H: 4.88(0.00)	G: 3.30(0.00) A: 4.54(0.00) H:-1.36(0.06)	G: 3.44(0.00) A: 6.22(0.00) H:-0.13(0.87)	G: 3.11(0.00) A: 5.30(0.00) H:-0.39(0.65)	G: 4.37(0.00) A: 5.61(0.00) H: 3.35(0.00)	G: 3.48(0.00) A: 5.19(0.00) H: 0.96(0.27)	G: 3.67(0.00) A: 5.66(0.00) H: 2.05(0.01)
20	n = 105		G: 0.68(0.32) A:-0.58(0.67) H:10.68(0.00)	G: 0.37(0.59) A:-0.43(0.71) H: 6.50(0.00)	G: 0.65(0.31) A:-0.57(0.59) H: 9.69(0.00)	G: 0.84(0.28) A: 2.28(0.09) H: 3.38(0.00)	G: 1.96(0.00) A:-0.10(0.93) H: 8.96(0.00)	G: 1.44(0.03) A: 1.44(0.21) H: 4.60(0.00)	G: 1.38(0.05) A: 0.86(0.51) H: 7.55(0.00)	G: 1.58(0.03) A: 1.40(0.26) H: 5.27(0.00)	G: 1.97(0.00) A: 1.73(0.14) H: 3.49(0.00)	G: 1.44(0.04) A: 1.25(0.37) H: 7.66(0.00)	G: 1.36(0.03) A: 1.55(0.19) H: 6.55(0.00)	G: 1.60(0.04) A: 2.70(0.04) H: 6.69(0.00)	G: 1.85(0.00) A: 1.84(0.12) H: 7.13(0.00)	G: 1.96(0.00) A: 2.69(0.02) H: 5.09(0.00)	G: 2.08(0.01) A: 3.16(0.01) H: 7.25(0.00)	G: 2.22(0.00) A: 3.01(0.02) H: 9.40(0.00)	G: 2.13(0.00) A: 1.82(0.13) H:11.72(0.00)	G: 2.54(0.00) A: 2.83(0.02) H: 5.52(0.00)	G: 2.68(0.00) A: 4.45(0.00) H: 6.70(0.00)	G: 2.37(0.00) A: 3.56(0.01) H: 6.91(0.00)	G: 3.16(0.00) A: 3.46(0.01) H:10.30(0.00)	G: 2.83(0.00) A: 3.54(0.00) H: 8.13(0.00)	G: 2.84(0.00) A: 3.96(0.00) H: 8.90(0.00)
125	n = 106	n = 105		G:-0.32(0.48) A: 0.10(0.91) H:-4.25(0.00)	G: 0.13(0.78) A: 0.41(0.58) H:-1.14(0.04)	G: 0.23(0.67) A: 2.86(0.00) H:-7.37(0.00)	G: 1.24(0.00) A: 0.48(0.57) H:-1.75(0.01)	G: 0.67(0.17) A: 1.75(0.05) H:-5.86(0.00)	G: 0.62(0.20) A: 1.35(0.09) H:-3.09(0.00)	G: 0.87(0.08) A: 1.93(0.02) H:-5.49(0.00)	G: 1.37(0.01) A: 2.35(0.01) H:-7.12(0.00)	G: 0.98(0.08) A: 2.52(0.00) H:-2.74(0.00)	G: 0.64(0.21) A: 2.31(0.01) H:-4.17(0.00)	G: 0.80(0.17) A: 3.14(0.00) H:-4.19(0.00)	G: 1.32(0.01) A: 2.45(0.01) H:-3.40(0.00)	G: 1.36(0.01) A: 3.49(0.00) H:-5.60(0.00)	G: 1.27(0.03) A: 3.59(0.00) H:-3.63(0.00)	G: 1.40(0.01) A: 3.16(0.00) H:-1.43(0.06)	G: 1.87(0.00) A: 3.08(0.00) H: 0.92(0.05)	G: 1.81(0.00) A: 3.31(0.00) H:-5.22(0.00)	G: 2.00(0.00) A: 5.03(0.00) H:-4.08(0.00)	G: 1.49(0.01) A: 3.76(0.00) H:-4.02(0.00)	G: 2.53(0.00) A: 4.36(0.00) H:-0.60(0.40)	G: 2.04(0.00) A: 4.05(0.00) H:-2.76(0.00)	G: 2.34(0.00) A: 4.66(0.00) H:-1.74(0.01)
25	n = 107	n = 106	n = 107		G: 0.61(0.14) A: 0.36(0.63) H: 3.26(0.00)	G: 0.47(0.36) A: 2.71(0.00) H:-3.13(0.00)	G: 1.63(0.00) A: 0.42(0.51) H: 2.46(0.00)	G: 1.14(0.02) A: 1.76(0.04) H:-1.36(0.05)	G: 0.99(0.01) A: 1.23(0.12) H: 0.96(0.16)	G: 1.20(0.00) A: 1.80(0.02) H:-1.26(0.11)	G: 1.57(0.00) A: 2.20(0.01) H:-2.93(0.00)	G: 1.38(0.00) A: 2.29(0.01) H: 1.54(0.03)	G: 1.07(0.01) A: 2.20(0.00) H: 0.12(0.84)	G: 1.19(0.03) A: 3.06(0.00) H: 0.06(0.94)	G: 1.64(0.00) A: 2.42(0.00) H: 0.72(0.24)	G: 1.59(0.00) A: 3.32(0.00) H:-1.42(0.06)	G: 1.66(0.00) A: 3.60(0.00) H: 0.62(0.47)	G: 1.67(0.00) A: 3.06(0.00) H: 2.61(0.00)	G: 2.19(0.00) A: 3.01(0.00) H: 5.44(0.00)	G: 2.19(0.00) A: 3.19(0.00) H:-0.96(0.19)	G: 2.31(0.00) A: 4.87(0.00) H: 0.20(0.81)	G: 1.92(0.00) A: 3.91(0.00) H: 0.26(0.76)	G: 3.40(0.00) A: 4.49(0.00) H: 4.23(0.00)	G: 2.40(0.00) A: 3.91(0.00) H: 1.44(0.08)	G: 2.58(0.00) A: 4.40(0.00) H: 2.34(0.00)
675	n = 106	n = 105	n = 106	n = 107		G:-0.16(0.74) A: 2.22(0.01) H:-6.53(0.00)	G: 1.11(0.01) A:-0.08(0.91) H:-0.63(0.15)	G: 0.70(0.12) A: 1.70(0.02) H:-4.70(0.00)	G: 0.34(0.43) A: 0.76(0.31) H:-2.26(0.00)	G: 0.59(0.20) A: 1.31(0.10) H:-4.53(0.00)	G: 1.15(0.02) A: 2.11(0.01) H:-6.20(0.00)	G: 0.75(0.12) A: 1.77(0.05) H:-1.88(0.00)	G: 0.54(0.25) A: 1.92(0.01) H:-3.18(0.00)	G: 0.58(0.23) A: 2.58(0.00) H:-3.25(0.00)	G: 1.10(0.02) A: 2.08(0.01) H:-2.60(0.00)	G: 1.08(0.02) A: 3.14(0.00) H:-4.77(0.00)	G: 1.06(0.03) A: 3.03(0.00) H:-2.66(0.00)	G: 1.05(0.02) A: 2.52(0.00) H:-0.58(0.35)	G: 1.59(0.00) A: 2.72(0.00) H: 2.09(0.00)	G: 1.76(0.00) A: 3.17(0.00) H:-4.28(0.00)	G: 1.71(0.00) A: 4.41(0.00) H:-3.15(0.00)	G: 1.15(0.02) A: 3.10(0.00) H:-2.68(0.00)	G: 2.47(0.00) A: 3.52(0.00) H: 0.31(0.63)	G: 1.81(0.00) A: 3.45(0.00) H:-1.86(0.01)	G: 2.07(0.00) A: 4.28(0.00) H:-0.93(0.18)
556	n = 105	n = 106	n = 105	n = 106	n = 105		G: 1.12(0.03) A:-2.38(0.01) H: 5.58(0.00)	G: 0.68(0.21) A:-0.94(0.36) H: 1.47(0.06)	G: 0.50(0.30) A:-1.49(0.04) H: 4.19(0.00)	G: 0.75(0.14) A:-0.88(0.34) H: 1.90(0.02)	G: 1.03(0.03) A:-0.73(0.45) H: 0.21(0.79)	G: 0.92(0.04) A:-0.48(0.51) H: 4.73(0.00)	G: 0.62(0.26) A:-0.51(0.59) H: 3.23(0.00)	G: 0.78(0.15) A: 0.42(0.68) H: 3.31(0.00)	G: 1.13(0.03) A:-0.44(0.63) H: 3.82(0.00)	G: 1.13(0.03) A: 0.61(0.50) H: 1.71(0.03)	G: 1.24(0.02) A: 0.88(0.39) H: 3.88(0.00)	G: 1.24(0.00) A: 0.37(0.61) H: 5.95(0.00)	G: 1.69(0.00) A: 0.46(0.55) H: 8.85(0.00)	G: 1.71(0.01) A: 0.55(0.58) H: 2.15(0.02)	G: 1.84(0.00) A: 2.16(0.00) H: 3.33(0.00)	G: 1.34(0.01) A: 1.05(0.24) H: 3.38(0.00)	G: 2.72(0.00) A: 1.71(0.13) H: 7.39(0.00)	G: 1.99(0.00) A: 1.26(0.22) H: 4.76(0.00)	G: 2.11(0.00) A: 1.73(0.09) H: 5.59(0.00)
5	n = 107	n = 106	n = 107	n = 108	n = 107	n = 106		G:-0.49(0.30) A: 1.27(0.16) H:-4.06(0.00)	G:-0.67(0.11) A: 0.79(0.28) H:-1.48(0.02)	G:-0.43(0.35) A: 1.38(0.08) H:-3.72(0.00)	G: 0.06(0.90) A: 1.86(0.04) H:-5.50(0.00)	G:-0.25(0.62) A: 1.99(0.03) H:-0.97(0.16)	G:-0.57(0.16) A: 1.77(0.02) H:-2.36(0.00)	G:-0.44(0.42) A: 2.63(0.01) H:-2.40(0.00)	G:-0.01(0.98) A: 1.96(0.01) H:-1.77(0.01)	G: 0.01(0.99) A: 2.99(0.01) H:-3.87(0.00)	G: 0.03(0.95) A: 3.08(0.00) H:-1.84(0.01)	G: 0.10(0.83) A: 2.63(0.00) H: 0.39(0.49)	G: 0.53(0.25) A: 2.86(0.00) H: 2.90(0.00)	G: 0.56(0.22) A: 2.77(0.01) H:-3.42(0.00)	G: 0.72(0.16) A: 4.54(0.00) H:-2.26(0.00)	G: 0.26(0.63) A: 3.46(0.00) H:-1.97(0.01)	G: 1.49(0.01) A: 4.29(0.00) H: 1.10(0.08)	G: 0.77(0.17) A: 3.49(0.00) H:-1.02(0.12)	G: 0.97(0.08) A: 4.02(0.00) H:-0.06(0.93)
50	n = 104	n = 103	n = 104	n = 105	n = 104	n = 103	n = 105		G:-0.06(0.91) A:-0.32(0.73) H: 2.54(0.00)	G: 0.12(0.80) A: 0.02(0.98) H: 0.36(0.61)	G: 0.55(0.23) A: 0.54(0.48) H:-1.00(0.14)	G: 0.26(0.54) A: 0.61(0.50) H: 3.14(0.00)	G:-0.01(0.98) A: 0.32(0.65) H: 1.81(0.01)	G: 0.06(0.92) A: 1.32(0.19) H: 1.49(0.06)	G: 0.57(0.17) A: 0.56(0.48) H: 2.40(0.00)	G: 0.53(0.32) A: 1.68(0.10) H: 0.06(0.93)	G: 0.52(0.36) A: 1.78(0.10) H: 2.03(0.01)	G: 0.60(0.21) A: 1.55(0.11) H: 4.16(0.00)	G: 0.94(0.06) A: 0.97(0.31) H: 6.87(0.00)	G: 1.16(0.03) A: 1.44(0.15) H: 0.81(0.23)	G: 1.15(0.03) A: 3.17(0.00) H: 1.66(0.03)	G: 0.66(0.22) A: 1.97(0.03) H: 1.62(0.03)	G: 1.94(0.00) A: 2.47(0.02) H: 5.55(0.00)	G: 1.20(0.03) A: 2.15(0.04) H: 2.80(0.00)	G: 1.37(0.02) A: 2.59(0.03) H: 3.78(0.00)
568	n = 107	n = 106	n = 106	n = 107	n = 106	n = 105	n = 107	n = 104		G: 0.22(0.58) A: 0.57(0.49) H:-2.29(0.00)	G: 0.52(0.25) A: 0.93(0.36) H:-4.14(0.00)	G: 0.37(0.37) A: 1.13(0.10) H: 0.44(0.53)	G: 0.08(0.85) A: 0.96(0.27) H:-0.89(0.22)	G: 0.24(0.66) A: 1.82(0.08) H:-0.94(0.24)	G: 0.63(0.14) A: 1.13(0.18) H:-0.36(0.57)	G: 0.60(0.25) A: 2.09(0.03) H:-2.49(0.00)	G: 0.72(0.21) A: 2.27(0.03) H:-0.39(0.64)	G: 0.76(0.08) A: 1.86(0.02) H: 1.77(0.00)	G: 1.26(0.00) A: 1.97(0.01) H: 4.27(0.00)	G: 1.21(0.02) A: 1.97(0.04) H:-1.95(0.01)	G: 1.32(0.01) A: 3.63(0.00) H:-0.82(0.21)	G: 1.01(0.07) A: 2.68(0.01) H:-0.60(0.48)	G: 2.25(0.00) A: 3.29(0.01) H: 2.70(0.00)	G: 1.39(0.01) A: 2.63(0.02) H: 0.37(0.62)	G: 1.58(0.01) A: 3.25(0.00) H: 1.33(0.08)
664	n = 107	n = 106	n = 107	n = 108	n = 107	n = 106	n = 108	n = 105	n = 107		G: 0.31(0.50) A: 0.30(0.74) H:-1.74(0.02)	G: 0.07(0.87) A: 0.33(0.71) H: 2.74(0.00)	G:-0.14(0.76) A: 0.38(0.61) H: 1.39(0.05)	G:-0.01(0.98) A: 1.25(0.18) H: 1.33(0.12)	G: 0.43(0.33) A: 0.57(0.50) H: 1.86(0.01)	G: 0.38(0.46) A: 1.50(0.16) H:-0.19(0.81)	G: 0.46(0.39) A: 1.70(0.07) H: 1.88(0.03)	G: 0.56(0.19) A: 1.45(0.08) H: 4.07(0.00)	G: 0.99(0.05) A: 1.37(0.12) H: 6.70(0.00)	G: 0.99(0.04) A: 1.39(0.14) H: 0.30(0.64)	G: 1.10(0.02) A: 3.06(0.00) H: 1.43(0.04)	G: 0.77(0.16) A: 2.22(0.01) H: 1.75(0.05)	G: 1.88(0.00) A: 2.60(0.01) H: 4.76(0.00)	G: 1.20(0.01) A: 2.11(0.02) H: 2.71(0.00)	G: 1.32(0.02) A: 2.47(0.02) H: 3.57(0.00)
38	n = 101	n = 102	n = 101	n = 102	n = 101	n = 102	n = 102	n = 99	n = 101	n = 102		G:-0.16(0.73) A: 0.19(0.85) H: 4.36(0.00)	G:-0.44(0.34) A:-0.00(1.00) H: 3.01(0.00)	G:-0.43(0.43) A: 0.83(0.41) H: 3.03(0.00)	G: 0.08(0.87) A: 0.19(0.83) H: 3.52(0.00)	G: 0.07(0.91) A: 1.22(0.26) H: 1.58(0.03)	G: 0.09(0.88) A: 1.35(0.20) H: 3.58(0.00)	G: 0.21(0.63) A: 1.21(0.18) H: 5.68(0.00)	G: 0.33(0.53) A: 0.32(0.74) H: 8.32(0.00)	G: 0.64(0.25) A: 1.30(0.21) H: 1.85(0.02)	G: 0.77(0.07) A: 2.76(0.00) H: 3.19(0.00)	G: 0.30(0.60) A: 1.90(0.06) H: 3.30(0.00)	G: 1.46(0.01) A: 1.87(0.09) H: 6.45(0.00)	G: 0.99(0.06) A: 1.87(0.06) H: 4.70(0.00)	G: 1.07(0.07) A: 2.35(0.04) H: 5.32(0.00)
658	n = 104	n = 103	n = 104	n = 105	n = 104	n = 103	n = 105	n = 102	n = 104	n = 105	n = 100		G:-0.26(0.56) A:-0.16(0.85) H:-1.42(0.02)	G: 0.04(0.94) A: 1.23(0.19) H:-1.29(0.10)	G: 0.29(0.50) A: 0.11(0.90) H:-0.87(0.18)	G: 0.43(0.37) A: 1.41(0.12) H:-2.71(0.00)	G: 0.54(0.31) A: 1.60(0.11) H:-0.73(0.37)	G: 0.39(0.40) A: 0.95(0.27) H: 1.19(0.07)	G: 0.86(0.09) A: 0.82(0.29) H: 3.95(0.00)	G: 0.93(0.09) A: 1.04(0.31) H:-2.41(0.00)	G: 0.91(0.06) A: 2.55(0.00) H:-1.27(0.12)	G: 0.53(0.34) A: 1.67(0.09) H:-1.10(0.25)	G: 1.79(0.00) A: 2.06(0.09) H: 2.55(0.00)	G: 1.32(0.01) A: 2.18(0.03) H: 0.13(0.88)	G: 1.42(0.01) A: 2.35(0.03) H: 1.01(0.23)
698	n = 106	n = 105	n = 106	n = 107	n = 106	n = 105	n = 107	n = 104	n = 106	n = 107	n = 101	n = 105		G: 0.18(0.73) A: 1.00(0.26) H:-0.06(0.95)	G: 0.56(0.16) A: 0.18(0.80) H: 0.54(0.28)	G: 0.52(0.33) A: 1.13(0.29) H:-1.49(0.07)	G: 0.64(0.23) A: 1.35(0.16) H: 0.50(0.59)	G: 0.60(0.19) A: 0.89(0.32) H: 2.55(0.00)	G: 1.10(0.04) A: 0.75(0.42) H: 5.32(0.00)	G: 1.13(0.02) A: 1.02(0.31) H:-1.09(0.09)	G: 1.24(0.02) A: 2.69(0.00) H: 0.20(0.82)	G: 0.86(0.10) A: 1.88(0.03) H: 0.33(0.73)	G: 1.98(0.00) A: 2.27(0.06) H: 3.93(0.00)	G: 1.41(0.01) A: 1.87(0.06) H: 1.34(0.15)	G: 1.47(0.01) A: 2.06(0.07) H: 2.27(0.01)
213	n = 107	n = 106	n = 107	n = 108	n = 107	n = 106	n = 108	n = 105	n = 107	n = 108	n = 102	n = 105	n = 107		G: 0.45(0.43) A:-0.64(0.52) H: 0.66(0.46)	G: 0.37(0.52) A: 0.20(0.85) H:-1.60(0.02)	G: 0.47(0.02) A: 0.45(0.31) H: 0.55(0.04)	G: 0.47(0.37) A:-0.11(0.91) H: 2.65(0.00)	G: 1.02(0.08) A: 0.01(0.99) H: 5.49(0.00)	G: 1.00(0.14) A: 0.14(0.91) H:-1.02(0.26)	G: 1.08(0.04) A: 1.75(0.08) H: 0.01(0.98)	G: 0.55(0.14) A: 0.66(0.34) H: 0.31(0.58)	G: 1.79(0.00) A: 1.07(0.36) H: 3.65(0.00)	G: 1.21(0.00) A: 0.86(0.14) H: 1.38(0.01)	G: 1.32(0.03) A: 1.22(0.29) H: 2.31(0.00)
651	n = 106	n = 105	n = 106	n = 107	n = 106	n = 105	n = 107	n = 104	n = 106	n = 107	n = 101	n = 104	n = 106	n = 107		G:-0.00(0.99) A: 1.00(0.37) H:-1.99(0.02)	G: 0.02(0.97) A: 1.10(0.26) H:-0.10(0.91)	G: 0.06(0.89) A: 0.73(0.43) H: 1.98(0.01)	G: 0.45(0.47) A: 0.61(0.52) H: 4.60(0.00)	G: 0.57(0.22) A: 0.83(0.39) H:-1.63(0.01)	G: 0.72(0.18) A: 2.56(0.01) H:-0.36(0.68)	G: 0.22(0.70) A: 1.55(0.10) H:-0.25(0.80)	G: 1.48(0.02) A: 2.11(0.08) H: 3.03(0.00)	G: 0.76(0.17) A: 1.50(0.14) H: 0.74(0.42)	G: 0.90(0.12) A: 1.88(0.13) H: 1.76(0.04)
26	n = 105	n = 106	n = 105	n = 106	n = 105	n = 106	n = 106	n = 103	n = 105	n = 106	n = 102	n = 103	n = 105	n = 106	n = 105		G: 0.12(0.84) A: 0.26(0.80) H: 2.17(0.00)	G: 0.01(0.98) A:-0.34(0.73) H: 4.13(0.00)	G: 0.42(0.46) A:-0.50(0.60) H: 6.97(0.00)	G: 0.58(0.28) A:-0.06(0.95) H: 0.44(0.59)	G: 0.72(0.21) A: 1.55(0.13) H: 1.62(0.03)	G: 0.46(0.43) A: 0.99(0.38) H: 2.30(0.00)	G: 1.48(0.01) A: 0.71(0.56) H: 5.67(0.00)	G: 0.86(0.13) A: 0.64(0.56) H: 3.05(0.00)	G: 0.98(0.01) A: 1.07(0.10) H: 3.83(0.00)
418	n = 107	n = 106	n = 107	n = 108	n = 107	n = 106	n = 108	n = 105	n = 107	n = 108	n = 102	n = 105	n = 107	n = 108	n = 107	n = 106		G:-0.02(0.97) A:-0.54(0.58) H: 2.09(0.00)	G: 0.55(0.33) A:-0.36(0.71) H: 4.88(0.00)	G: 0.53(0.42) A:-0.31(0.79) H:-1.57(0.09)	G: 0.60(0.27) A: 1.29(0.20) H:-0.55(0.38)	G: 0.09(0.78) A: 0.22(0.71) H:-0.19(0.71)	G: 1.33(0.04) A: 0.60(0.61) H: 3.18(0.00)	G: 0.74(0.08) A: 0.41(0.53) H: 0.63(0.10)	G: 0.84(0.17) A: 0.85(0.46) H: 1.69(0.03)
671	n = 106	n = 105	n = 106	n = 107	n = 106	n = 105	n = 107	n = 104	n = 106	n = 107	n = 101	n = 104	n = 106	n = 107	n = 106	n = 105	n = 107		G: 0.49(0.36) A: 0.09(0.92) H: 2.50(0.00)	G: 0.51(0.34) A: 0.19(0.85) H:-3.60(0.00)	G: 0.49(0.25) A: 1.50(0.06) H:-2.66(0.00)	G: 0.20(0.67) A: 0.88(0.29) H:-2.28(0.01)	G: 1.08(0.07) A: 0.60(0.57) H: 0.68(0.31)	G: 0.65(0.23) A: 0.65(0.53) H:-1.37(0.04)	G: 0.74(0.16) A: 1.04(0.31) H:-0.59(0.38)
307	n = 103	n = 102	n = 104	n = 104	n = 104	n = 102	n = 104	n = 101	n = 103	n = 104	n = 98	n = 102	n = 104	n = 104	n = 103	n = 102	n = 104	n = 103		G: 0.16(0.76) A: 0.52(0.58) H:-6.50(0.00)	G: 0.16(0.78) A: 1.86(0.03) H:-5.38(0.00)	G:-0.33(0.54) A: 0.77(0.38) H:-4.95(0.00)	G: 0.78(0.25) A: 1.02(0.42) H:-1.59(0.01)	G: 0.25(0.67) A: 0.93(0.39) H:-4.02(0.00)	G: 0.53(0.38) A: 1.55(0.15) H:-3.18(0.00)
113	n = 107	n = 106	n = 107	n = 108	n = 107	n = 106	n = 108	n = 105	n = 107	n = 108	n = 102	n = 105	n = 107	n = 108	n = 107	n = 106	n = 108	n = 107	n = 104		G: 0.14(0.83) A: 1.61(0.15) H: 1.18(0.18)	G:-0.27(0.67) A: 0.82(0.46) H: 1.35(0.19)	G: 0.52(0.38) A: 0.69(0.53) H: 4.38(0.00)	G: 0.22(0.73) A: 0.73(0.57) H: 2.40(0.01)	G: 0.36(0.54) A: 1.18(0.27) H: 3.36(0.00)
105	n = 105	n = 106	n = 105	n = 106	n = 105	n = 106	n = 106	n = 103	n = 105	n = 106	n = 102	n = 103	n = 105	n = 106	n = 105	n = 106	n = 106	n = 105	n = 102	n = 106		G:-0.37(0.49) A:-0.84(0.36) H: 0.39(0.56)	G: 0.69(0.10) A:-0.50(0.65) H: 3.90(0.00)	G: 0.15(0.78) A:-0.91(0.34) H: 1.43(0.02)	G: 0.23(0.67) A:-0.45(0.65) H: 2.21(0.00)
214	n = 100	n = 99	n = 100	n = 101	n = 100	n = 99	n = 101	n = 99	n = 100	n = 101	n = 95	V = 98	n = 100	n = 101	n = 100	n = 99	n = 101	n = 100	n = 97	n = 101	n = 99		G: 1.21(0.06) A: 0.54(0.63) H: 3.17(0.00)	G: 0.77(0.16) A: 0.47(0.59) H: 0.97(0.18)	G: 0.54(0.41) A: 0.21(0.87) H: 1.93(0.02)
710	n = 92	n = 91	n = 92	n = 93	n = 93	n = 91	n = 93	n = 91	n = 92	n = 93	n = 89	n = 91	n = 92	n = 93	n = 92	n = 91	n = 93	n = 92	n = 91	n = 93	n = 91	n = 87		G:-0.51(0.39) A:-0.12(0.91) H:-2.40(0.00)	G:-0.36(0.54) A: 0.43(0.72) H:-1.69(0.02)
212	n = 107	n = 106	n = 107	n = 108	n = 107	n = 106	n = 108	n = 105	n = 107	n = 108	n = 102	n = 105	n = 107	n = 108	n = 107	n = 106	n = 108	n = 107	n = 104	n = 108	n = 106	n = 101	n = 93		G: 0.11(0.85) A: 0.37(0.74) H: 0.82(0.26)
27	n = 104	n = 105	n = 104	n = 105	n = 104	n = 105	n = 105	n = 102	n = 104	n = 105	n = 101	n = 102	n = 104	n = 105	n = 104	n = 105	n = 105	n = 104	n = 101	n = 105	n = 105	n = 98	n = 90	n = 105	

The results of paired Student's *t*-test on common targets are reported in the form of the mean of the differences in GDT-HA (G), AL-0 (A), and HBscore (H) values along with the associated *P*-values in parentheses. Statistically significant differences between groups (*P*-values < 0.05) are shaded in gray. Number of common targets in the comparison are reported in the lower left half of the table.

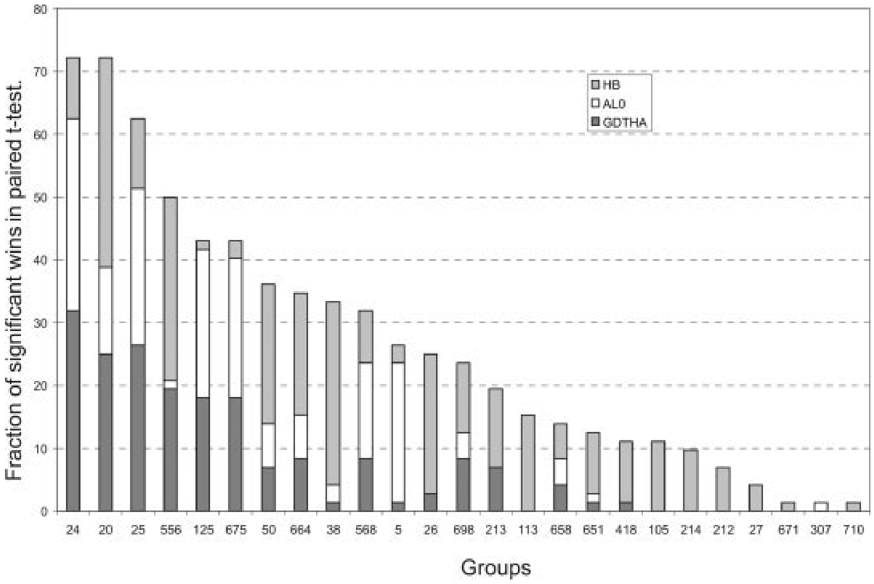


Figure 5

Head-to-head comparison for the top 25 groups showing the fraction of statistically significant wins (Student's t-test P-*value* < *0.05) on common targets.*

that were worse than average, i.e. had negative *z*-scores, and those flagged as physically impossible were assigned *z*-scores of 0. Setting the *z*-scores of models of average or worse quality to zero ensures that the submitting group is not excessively penalized in the overall scoring, thereby encouraging the application and development of innovative and somewhat riskier methods. For each group, we calculated mean *z*-scores for GDT-HA and AL0, as well as a combined mixed *z*-score as the average of both. Table II summarizes the results for each predictor group: The number of models assessed, the mean values, and mean *z*-scores for GDT-HA and AL0, and the combined GDT-HA/AL0 *z*-score. Prediction groups registered as servers are marked with "s." From this list, we selected the 25 highest scoring groups based on the combined *z*-score for a more detailed assessment.

The results of the top 25 groups were compared by direct head-to-head comparison on common targets using paired Student's *t*-tests, as introduced in CASP5.[3] Note that these comparisons were based on the raw scores of GDT-HA, AL0, and HBscore values for each prediction (Table III). For models worse than average (i.e. negative *z*-scores) the raw scores were set to target average (corresponding to $z = 0$). HBscore values for all assessment units for the top 25 groups are provided as Supplementary Materials (Table S-I). Models flagged as physically impossible were omitted from the head-to-head comparison. Finally, the number of statistically significant wins over the other groups (Student's *t*-test *P*-value < 0.05) on common targets was calculated for all three measures, and summed up for each group. Figure 5 shows the fraction of statistically significant wins in the head-to-head comparison for the top 25 groups. The six groups ranked top according to the combined *z*-scores were the same ranked highest in the head-to-head comparison: 24 (Zhang), 20 (Baker), 25 (Zhang-Server), 556 (LEE), 125 (TASSER), and the meta-predictor group 675 (Fams-ace). Groups 24 and 20 produced on average models of higher quality. Remarkably, the automated protein modeling server of group 25 generated on average models of nearly comparable quality to the two leading manual predictor groups. In contrast to earlier CASP experiments, the methods of all top scoring groups were highly automated computational approaches. This reflects on one side the results of ongoing method development in recent years, but partly may also be a sign of the time constraints of manual groups during the modeling season caused by the relatively large number of prediction targets in CASP7.

During the meeting in Asilomar, it became clear that the top scoring methods—although producing models of comparable backbone quality—differ significantly in their algorithmic approaches, computational requirements, modeling of side chain packing, and atomic interactions. This indicates possible directions for further development of the individual methods.

Improvement over the best single template

TBM procedures rely on the detection of and correct alignment to homologous template structures. Consequently, the resulting structure models are generally

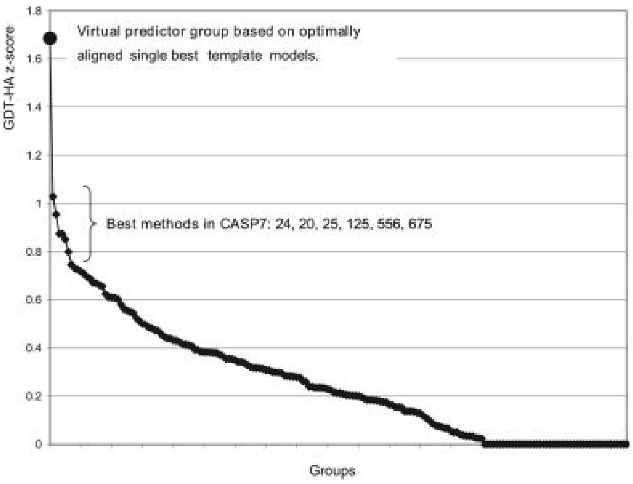


Figure 6

GDT-HA z-scores for all CASP7 methods in comparison with a "virtual predictor group" based on optimally aligned single best template models.

closer to the template than to the target. Recently, several methods have been claimed by their developers to be able to improve over the template. We have therefore assessed whether the predictions submitted to CASP7 showed improvement over the best available single template structure. For this purpose, the PDB was searched for suitable templates available at the end of the prediction period for each target, using LGA[10] and Mammoth[11] as described elsewhere.[7] Based on the structural alignments generated by a 4 Å LGA sequence-independent superposition, we generated pseudo-predictions by copying the backbone coordinates of the templates. No coordinates were assigned to unaligned residues in insertions and deletions. A "virtual predictor group" using these pseudo-predictions would on average outperform all other methods by far, as shown in Figure 6. However, for individual targets, some groups succeeded in building models better than the pseudo-prediction based on the single best template. The best group in this respect (24, Zhang) managed to achieve a higher GDT-TS score than the virtual group in more than half the assessment units and a higher GDT-HA score in approximately one-third of cases. Figure 7 illustrates the fraction of targets for which each group predicted more accurate models than the best single template (plotted on the positive *y*-axis) and targets predicted worse than the best single template (negative *y*-axis).

For a small number of targets, the submitted predictions were significantly better than the best single template model. The most remarkable example was target T0283, where the best prediction showed an improvement of 20.4 GDT-HA units. Several different effects may account for the observed improvements over single template pseudo-predictions, e.g. including information from multiple templates, modeling of insertions, deletions, structurally diverse regions, and refinement to improve the overall quality of the model. A more detailed discussion is provided elsewhere in this issue.[12] Overall however, most of the observed improvements are rather small. In Figure 7, predictions with differences of less than 2.0 GDT-HA units, between the model and the best template are shaded in black. For the majority of targets, the observed differences are still small compared to the overall modeling error—too small to make a significant difference in most biological applications.

Comparison between CASP6 and CASP7

Comparisons between different rounds of CASP are difficult as targets pose very diverse challenges to the predictors, and the modeling difficulty can only be roughly estimated by a combination of various parameters.[1,13] Overall, no large improvement in general model accuracy has been observed between CASP6 and CASP7 when comparing average GDT-TS values as a function of modeling difficulty (data not shown). To assess more subtle improvements, we performed a comparison between CASP6 and CASP7 by evaluating the ability of the methods to improve their models over the best available single template. We applied the "best template model" procedure described above to the CASP6 targets classified as comparative modeling or homologous fold recognition targets (CM and FR/H) based on the best template structures used for the CASP6 assessment.[2,14] Figure 8 shows the average fraction of predictions that boast an improvement of more than 0.0, 0.5, 1.0, 2.0, 4.0, and 8.0 GDT-HA units over the template, for the best ten groups in CASP6 and

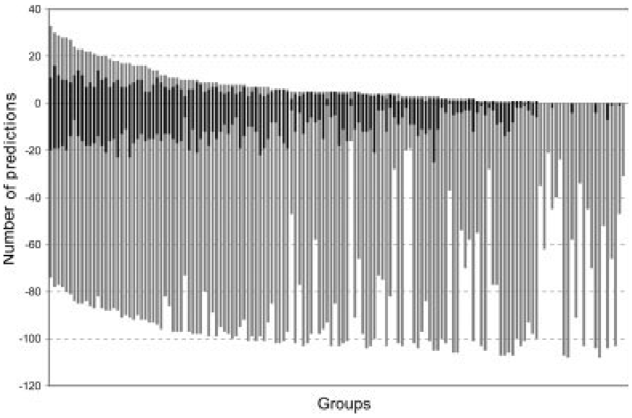


Figure 7

Performance relative to the single best template. For each group, the number of targets predicted more accurately than expected for a model based on the best single template is plotted on the positive y-axis, targets predicted worse are plotted negative. Small differences less than 2.0 GDT-HA units are shaded in black.

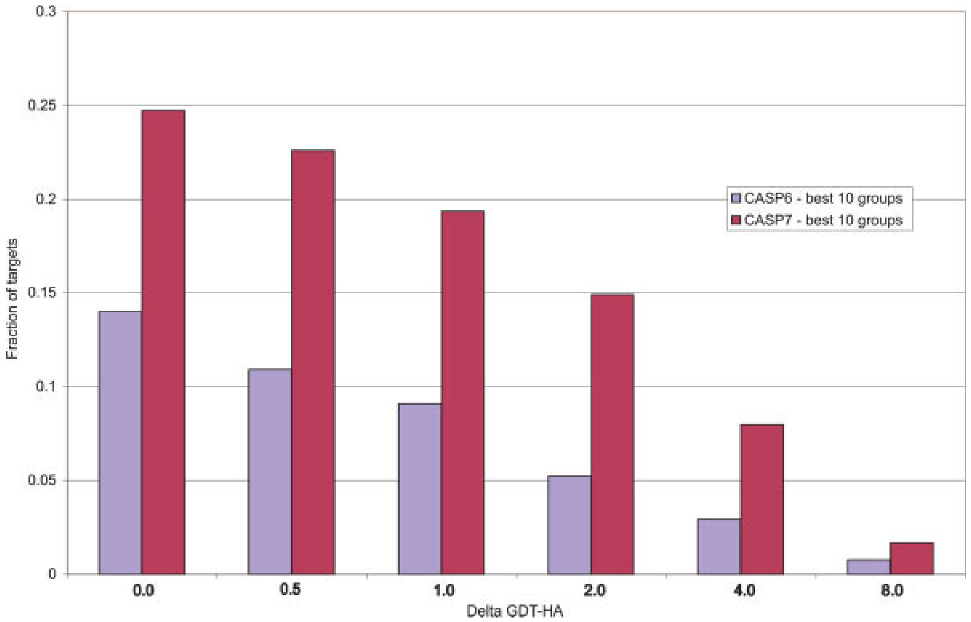


Figure 8

Performance relative to the single best template in comparison between CASP6 and CASP7. The average fractions of predictions, which achieved an improvement compared to the best template model of at least 0.0, 0.5, 1.0, 2.0, 4.0, and 8.0 GDT-HA units are shown for the best 10 groups according to this criterion in both CASP6 (blue) and CASP7 (purple). [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

CASP7. Assuming that both CASP6 and CASP7 experiments were of comparable difficulty—which is consistent with a Wilcoxon Rank Sum Test ($P = 0.31$) of the target difficulty based on the scale used in CASP6[2]—we observe that a higher number of groups were able to generate a larger fraction of predictions showing improvement over the best available templates in CASP7.

The observed improvements can be attributed to several factors. Methods that combine multiple template information have made substantial progress in recent years, and none of the top ranked groups in CASP7 was using single template approaches for model building. Multiple template and fragment-based methods can also make effective use of the increased coverage of structure space by structural genomics and individual structure elucidation efforts. Additionally, methods developed for refining template-free models may account for the observed improvement in cases with only limited template structure information. One of the most astonishing examples was target T0283, where two predictor groups have submitted models that were of significantly better quality than the remainder (Fig. 9): Group 20 (Baker) with a GDT-HA of 59.3, and AL0 of 77.3, and group 13 (Jones) with a GDT-HA of 45.1 and AL0 of 62.9. The best available template structure was the archaeal ammonium transporter Amt-1 from *Archaeoglobus fulgidus* (PDB: 2b2j) with an RMSD of 2.54 Å, sharing only 6.7% sequence identity with the target. The best models have significantly lower RMSD values of 1.78 Å and 2.37 Å, respectively. Both groups describe their methods as fragment assembly approaches with a subsequent refinement step. As illustrated in Figure 9(d), the prediction by group 20 is characterized by remarkably accurate local interactions and packing of side chains for a prediction of such low similarity to the closest template and indicates a successful atomic refinement of the model.

There is no optimal method for generating pseudo-predictions for this analysis. Different parameters can be used for identifying structural templates for a target protein, generating structural superpositions, deriving structural alignments, and building the pseudo-models. Pseudo-predictions built using different protocols can differ and data derived from them may vary. Nevertheless, the improvement observed from CASP6 to CASP7 as shown in Figure 8 is stable and largely independent of parameter choice. It should be noted that in most cases the amount of improvement observed for individual targets is relatively small compared to the overall modeling error. This may explain why improvement is only observed by using a “best template model” as internal reference point for each target, while none is detected when using the classical overall difficulty scale. A detailed discussion on progress over previous CASP experiments is provided elsewhere in this issue.[12]

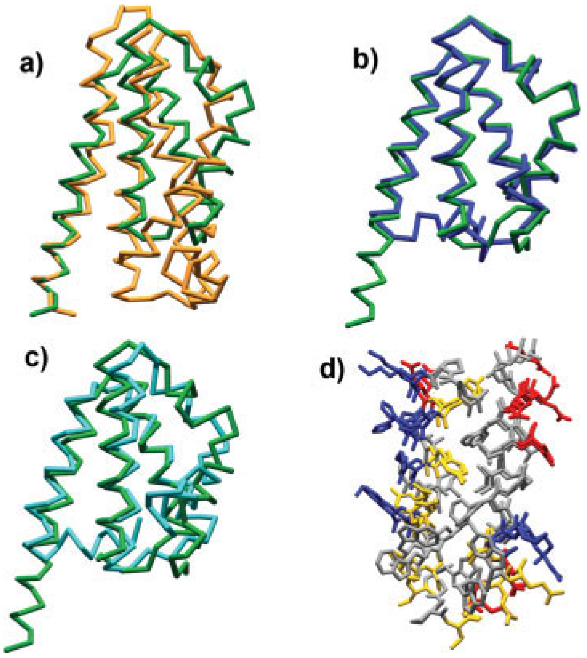


Figure 9

Examples of model quality: (a) Superposition of target structure T0283 (green) and best template (PDB: 2b2j, orange), with an RMSD of 2.54 Å sharing 6.7% sequence identity. The models submitted by two groups were of significantly better quality than the remainder: Group 20 (Baker) with a GDT-HA of 59.3, and AL0 of 77.3 shown in (b) in dark blue. The model by group 13 (Jones) achieved a GDT-HA of 45.1 and AL0 of 62.9 and is shown in (c) in light blue. (d) Detailed view of the side chain packing of the target structure and the model submitted by group 20.

Accuracy of binding site predictions

Active sites or cofactor binding sites in protein models are of great interest for biologists using protein structures or models in their daily work. For several prediction targets, the CASP organizers released the target sequence together with information about a ligand bound in the target structure, which should enable the predictors to model functionally important residues in the binding site more accurately. For the assessment of this aspect, we superposed the models onto the target structure based on only the C_α positions of residues interacting with the ligand in the crystal structure. We evaluated the quality of the modeled binding site using an atomic contact score (ACS), which considers interactions between the nonhydrogen atoms of the protein and the ligand [Eq. (1)]:

$$\mathrm{ACS} = \frac{\sum_{i,k} (\mathrm{Cont}_{i,k}^{\mathrm{target}} \cdot \mathrm{Cont}_{i,k}^{\mathrm{model}}) - \sum_{i,k} \mathrm{Clash}_{i,k}^{\mathrm{model}}}{\sum_{i,k} \mathrm{Cont}_{i,k}^{\mathrm{target}}} \qquad (1)$$

with

$$\mathrm{Cont}_{i,k} = \begin{cases} 1 & \text{if } 2.0\text{Å} \le r_{i,k} \le 4.0\text{Å} \\ 0 & \text{otherwise} \end{cases},$$

$$\mathrm{Clash}_{i,k} = \begin{cases} 1 & \text{if } r_{i,k} \le 1.5\text{Å} \\ 0 & \text{otherwise} \end{cases}.$$

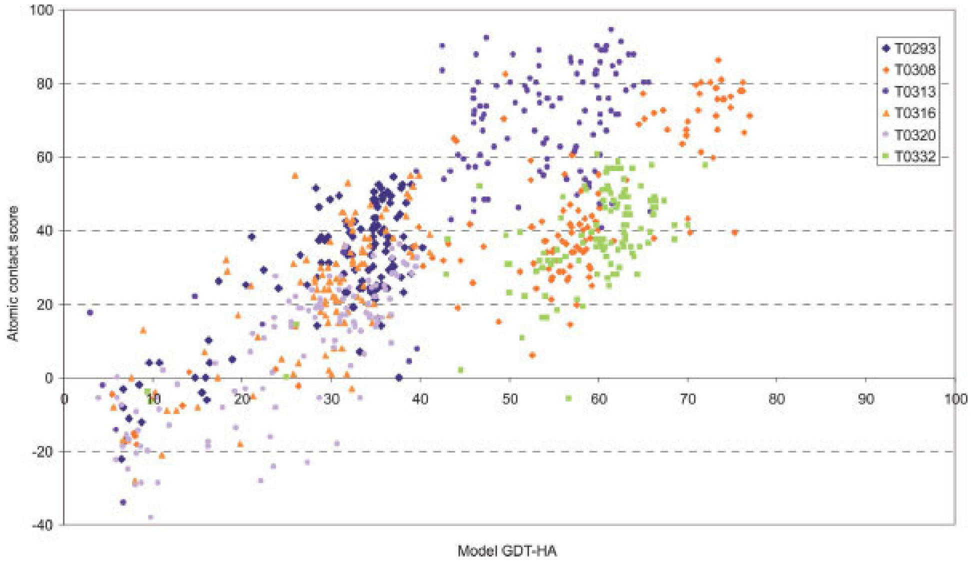


Figure 10

Accuracy of cofactor binding site predictions of six TBM targets. The fraction of correctly modeled atomic interactions in the binding sites (see text) is plotted against the overall model accuracy GDT-HA.

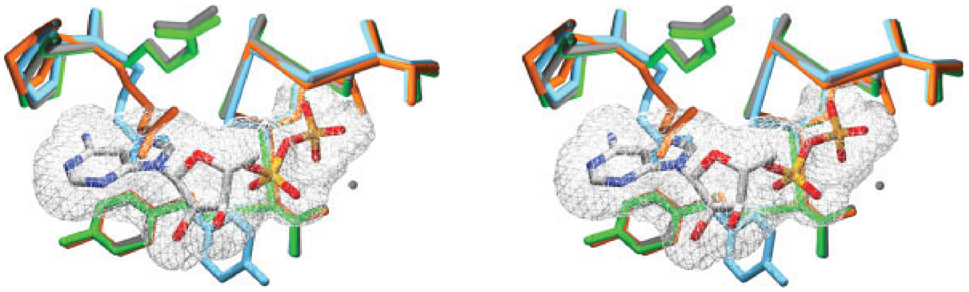


Figure 11

Superposition of experimental and predicted ADP binding sites of target T0313, a human KIFC3 motor domain (stereo view). The experimental structure with the ADP ligand and its solvent accessible surface are shown in gray, the best prediction by group 186 in green, and predictions by groups 20 in orange and 24 in light blue.

We evaluated the fraction of correctly modeled atomic contacts in the predicted binding sites by enumerating specific contacts between the protein atoms (i) and the ligand atoms (*k*) using NCONT.[15] The score in Eq. (1) penalizes interactions, which are predicted shorter than 1.5 Å by classifying them as clashes. Figure 10 illustrates the percentage of correctly modeled contacts in six prediction targets with bound cofactors: *S*-adenosylhomocysteine in T0293 and T0332, GTP in T0308, ADP in T0313, *S*-adenosylmethionine in T0316, and FAD in T0320. Except for T0316, information about the bound ligand was provided along with the prediction target sequence. We would like to emphasize that this analysis of binding site accuracy is based on six examples and therefore has limited statistical power. It only allows for a qualitative description, but not a quantitative assessment of the ability of individual groups to accurately model cofactor binding sites.

The ADP binding site of T0313, a human KIFC3 motor domain, is formed by 12 residues, nine of which are shown in Figure 11. The experimental structure with the ADP ligand and its solvent accessible surface are shown in gray. The best predictions (Fig. 11, green) reproduce the atomic interactions formed by 18 backbone and 24 side chain atoms very well. Models using human mitotic spindle kinesin Eg5 (PDB: 1ii6, chain A) with bound ADP as single template reproduce both main chain and side chain geometry successfully. However, numerous models with accurate backbone geometry were submitted, which fail to form an intact ADP binding site. Figure 11 shows two examples in which the side chain of Arg 9 protrudes into the ADP binding site (orange and light blue), and the stacking interaction by Tyr 92 is modeled incorrectly (light blue).

Overall, the accuracy of the predicted binding sites varies significantly between target structures. Compared to the other examples, T0313 represents a relatively simple modeling task as the alignment is unambiguous. For T0313, the best groups manage to reproduce more than 90% of the ligand–protein interactions, while even in the best predictions for T0320 this fraction is lower than 40% (Fig. 10). While there is a general trend for models with inaccurate backbone geometry to have incorrectly modeled binding sites, it does not hold for models with a GDT-HA above 40. In fact, for T0313, the best five binding site models vary in GDT-HA from 42.6 to 62.6. Therefore, global C_α-based measures such as GDT-HA cannot be used as the only criterion to indicate biological relevance of a model. Also, on average no significant difference in prediction accuracy of the binding site was observed between targets for which the bound cofactor was announced with the prediction target, and T0316 for which this information was not directly available to the predictors. In conclusion, it appears that modeling biologically relevant features of the target proteins as accurately as possible has not received the same level of attention by all predictor groups in CASP7.

Model quality estimates

The practical application of protein models strongly depends on their quality. However, at the time of model generation, the correct answer is unknown and the accuracy of the model must therefore be estimated beforehand. We have assessed a posteriori the ability of individual CASP7 modeling groups to assign realistic error estimates to their predictions. For all targets in the TBM category, we calculated the model error for each predicted amino acid residue as the Cartesian distance between the model C_α coordinates and the experimental target structure in a global superposition with LGA[10] using a 4 Å cut-off. For each participating group (i.e. groups who submitted at least two different values in the *B*-factor column for more than 10 targets), the accuracy of the error estimates ("Model *B*-factor") was analyzed using a log-linear correlation between the estimates and the real error (Fig. 12). Additionally, the results of a random model predictor were added to the analysis for comparison. To compile the data of this null model, "Model *B*-factor" values were

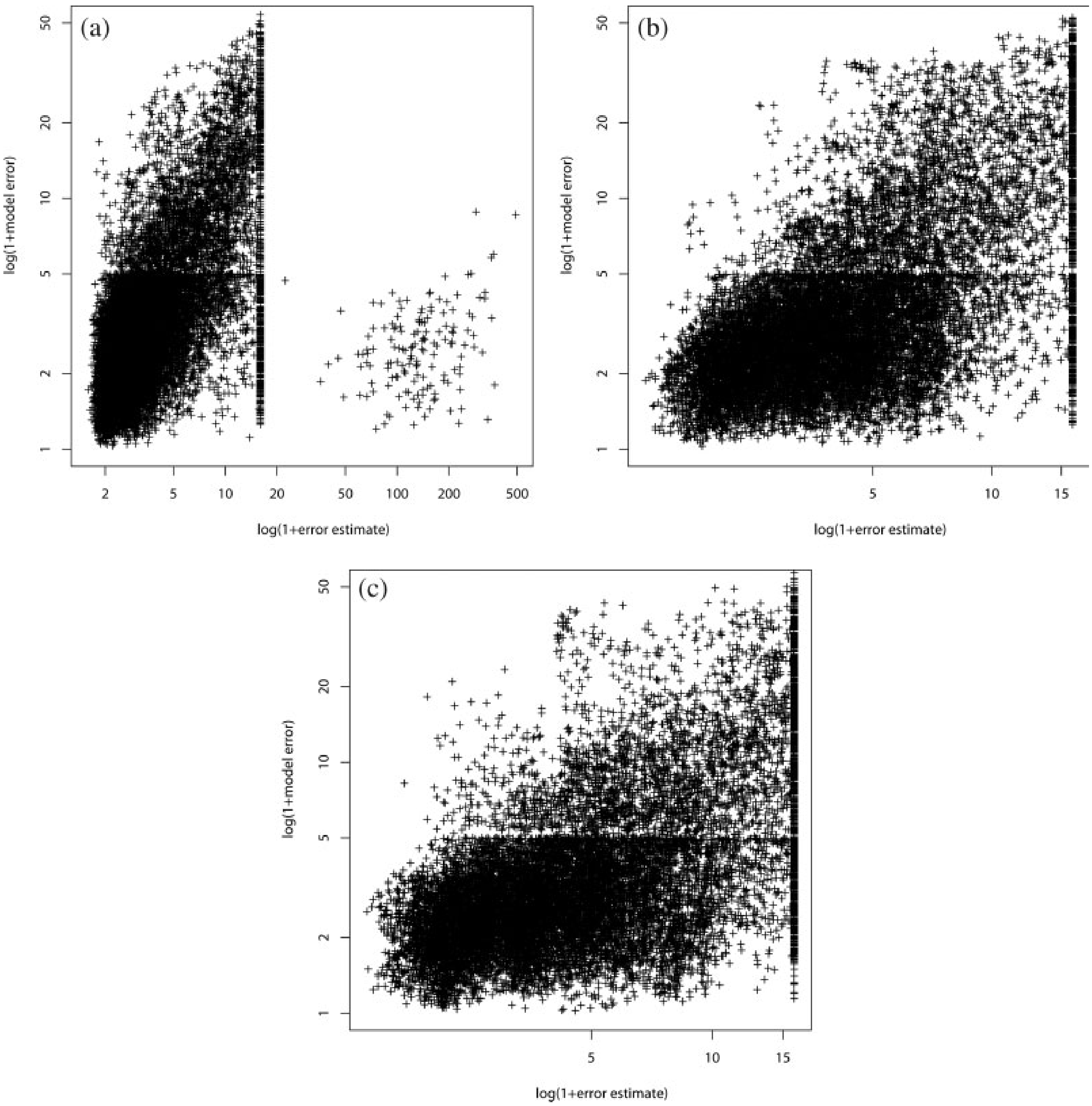


Figure 12

Log-linear correlation between the estimated model error ("Model B*-Factor") and the actual error of the model for (a) group 50, (b) group 46, and (c) group 47.*

randomly chosen from the list of model-target distances. Since linear correlation analysis is sensitive to outliers, the prediction results were additionally analyzed using receiver operator characteristic curves (ROC) (Fig. 13). We classified a residue as correctly modeled if its C_α position error is less than 3.8 Å and incorrectly if its C_α position error is greater than or equal to 3.8 Å. For each group, the "Model *B*-factor" of the predictions were reranked between 0 and 1 and the enrichment of correctly identified model errors was plotted as the false positive rate (FPR) versus the true positive rate (TPR) by varying the discrimination threshold between 0 and 1. The TPR is defined as the number of true positives (TP, the number of correctly identified model errors) over the total number of model errors (positives, *P*), and the FPR the number of false positives (FP, identified as errors in the model, but in reality modeled correctly) over the total number of correctly modeled residues (negatives, *N*). The

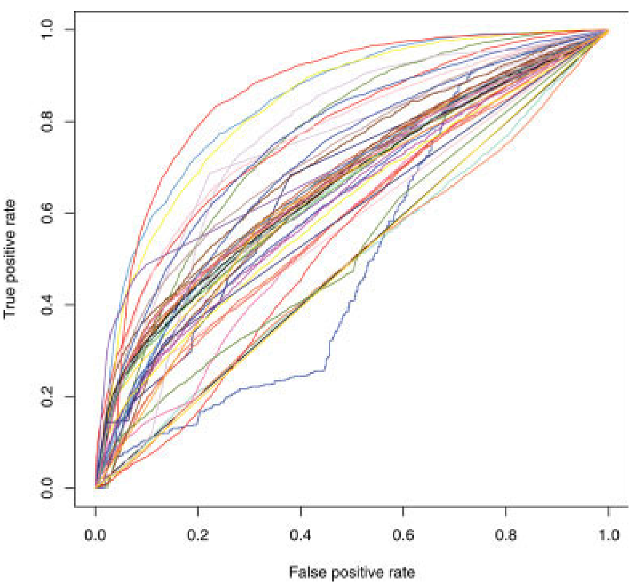


Figure 13

ROC curves analyzing the accuracy of the estimated model error ("Model B-Factor") to correctly identify incorrect residues defined as C_α error greater than or equal to 3.8 Å. Groups with highest ROC AUC are 50 (red), 46 (light blue), and 47 (yellow).

area under the curve (AUC) is used as a measure for the accuracy in correctly identifying model errors.

The results of the correlation analysis and the ROC curves for each of the 60 predictor groups, which submitted model error estimates, are listed in Table IV sorted by decreasing AUC. The best-performing methods according to ROC curves by Eloffson and coworkers (50, SBC; 46, Pcons6; 47, Pmodeller6) display also the highest linear correlation coefficients ranging from 0.530 to 0.620. Although many groups were using their own metric, the best groups provided model error estimates based on an absolute metric in Ångstrom as specified by the CASP format. In summary, although only 32% of the groups have provided confidence values for their predictions, the results of the present CASP experiment are encouraging. Regarding the type of methods used, it appears that consensus-based approaches[16] using server predictions submitted to CASP7 as input outperformed other approaches based solely on physics, statistical measures, and traditional model quality estimation programs (MQEP).[17]

CONCLUSIONS

From the perspective of a method developer, the aim of the assessment of template-based protein structure prediction in CASP is to establish the state of the art in the field, identify progress, and pinpoint bottlenecks in areas where further research is required. From the perspective of the life science community, template-based protein structure prediction and modeling has come of age and is widely used today as a scientific research tool. Therefore, an increasingly important aspect of the assessment is also to evaluate to what extent today's prediction methods meet the accuracy requirements of different scientific applications.

In CASP7, we could base our evaluation on a large number of predictions, which provided a solid basis to

Table IV
Assessment of Model Error Estimates

No.	Groups	ROC (AUC)	Correlation (*r*)	No.	Groups	ROC (AUC)	Correlation (*r*)	No.	Groups	ROC (AUC)	Correlation (*r*)
1	050	0.858	0.528	21	248	0.665	0.288	41	401	0.608	0.184
2	046	0.844	0.616	22	347	0.662	0.259	42	698	0.600	0.154
3	047	0.832	0.573	23	658	0.660	0.270	43	416	0.596	0.250
4	005	0.771	0.526	24	413	0.660	0.263	44	137	0.584	0.124
5	214	0.765	0.434	25	609	0.659	0.251	45	136	0.583	0.149
6	275	0.762	0.445	26	415	0.656	0.260	46	654	0.583	0.098
7	038	0.753	0.287	27	728	0.655	0.243	47	453	0.583	0.145
8	368	0.743	0.396	28	414	0.653	0.254	48	020	0.576	0.234
9	004	0.742	0.492	29	074	0.651	0.263	49	083	0.570	0.088
10	060	0.722	0.393	30	087	0.649	0.264	50	261	0.541	0.084
11	297	0.705	0.491	31	267	0.649	0.227	51	659	0.534	0.002
12	025	0.704	0.326	32	490	0.643	0.105	52	338	0.527	0.027
13	274	0.699	0.299	33	021	0.641	0.224	53	013	0.503	0.074
14	212	0.684	0.322	34	337	0.635	0.206	54	063	0.498	0.013
15	103	0.676	0.299	35	494	0.634	0.377	55	420	0.495	0.008
16	677	0.676	0.299	36	427	0.629	0.296	56	474	0.494	0.020
17	213	0.672	0.280	37	105	0.626	0.200	57	071	0.493	0.009
18	092	0.671	0.283	38	651	0.626	0.226	58	483	0.491	0.010
19	091	0.669	0.253	39	319	0.622	0.187	59	203	0.488	0.034
20	418	0.665	0.254	40	024	0.621	0.172	60	614	0.487	0.041

For the 60 groups providing model error estimates for their predictions, the accuracy of residue-based error estimates was assessed by ROC and log-linear correlation using differences between the individual model and the target structure as reference.

assess the differences between the participating prediction methods. We have adapted the numerical assessment criteria to account for the scientific progress in the field of TBM: we applied a global distance test with stricter cutoff values (GDT-HA 0.5, 1, 2, 4 Å) to focus more on the finer details of the predictions. Additionally, to complement GDT and AL0 scores, which are based solely on global superpositions of C_α atoms, we introduced HBscore as local atomic measure evaluating how well hydrogen bond interactions in the target structure are reproduced in the model. Local atomic measures such as HBscore can discriminate between predictions with otherwise similar C_α structures. We have observed significant differences in the accuracy of modeling atomic interactions of backbone hydrogen bonds and side chain packing, indicating areas for further improvement for many of the participating methods.

Overall, the top scoring groups relied on highly automated computational approaches with limited manual intervention. Remarkably, one automated modeling server produced models of nearly comparable quality to the two leading manual predictor groups. We analyzed the ability of different methods to generate predictions that improve over a model based on a single best template structure. Compared to CASP6, a higher number of methods were able to achieve improvement over the best template. It appears that several methods make effective use of multiple template structures. In some cases with limited template information, the observed improvement over template can be attributed to successful application of fragment based modeling or model refinement methods. Although the observed improvement over the best template model is a promising step in the right direction, it is mostly very limited and often cannot be considered as biologically relevant. The fact that no group would outperform a "virtual predictor" submitting models based on the single best template for each target indicates that template identification and alignment are by no means solved problems and constitute a major bottleneck in TBM, besides the challenging question of model refinement.

For regions of the protein, which are of functional importance, such as active sites or ligand binding pockets, accurate reproduction of local interactions such as hydrogen bonds and side chain conformations is essential. For six CASP7 target structures with bound cofactors, we observed considerable differences between models in terms of local model accuracy, even when their C_α structures are similar. Since the accuracy of functional regions is a limiting factor for the scientific usefulness of the predicted structure, improvements in this area would be a big benefit for the life science community.

According to Henry A. Bent*, "... a model must be wrong, in some respects—else it would be the thing itself. The trick is to see ... where it's right."[18] In other words, accurate estimates of the errors of a model are an essential component of any predictive method—protein structure prediction not being an exception. Therefore, we have (for the first time in CASP) evaluated the accuracy of the expected model errors provided by the predictors for their models. Unfortunately, only one-third of all predictor groups provided these for their predictions. Clearly, consensus-based methods gave the most accurate error estimates. From our point of view, confidence measures are an essential part of a prediction method both from a methods development and practical application perspective, and should therefore be an integral component of future assessments.

*Deliberately quoted out of context.

ACKNOWLEDGMENTS

We are grateful to the Prediction Center, especially Andriy Krystafovych, for providing data for the analysis and fast and profound support. We wish to thank our fellow assessors Randy Read and Neil Clarke for the constructive collaboration, and Ernest Feytmans for advice on the statistical analysis. We are grateful to the CASP6 assessors B. K. Lee, Alfonso Valencia, and Roland Dunbrack for valuable advice on the CASP assessment, and to Michael Tress for providing template information for CASP 6 and 7 targets. We gratefully acknowledge support of our group by the Swiss Institute of Bioinformatics. We thank all experimental groups and participating predictors, without whom CASP would not be possible. Last but not least, we thank John Moult, Burkhard Rost, Tim Hubbard, and especially Anna Tramontano for continuous encouragement and valuable discussions.

REFERENCES

1. Kryshtafovych A, Venclovas C, Fidelis K, Moult J. Progress over the first decade of CASP experiments. Proteins 2005;61(Suppl 7):225–236.
2. Tress M, Ezkurdia I, Grana O, Lopez G, Valencia A. Assessment of predictions submitted for the CASP6 comparative modeling category. Proteins 2005;61(Suppl 7):27–45.
3. Tramontano A, Morea V. Assessment of homology-based predictions in CASP5. Proteins 2003;53(Suppl 6):352–368.
4. Wang G, Jin Y, Dunbrack RL, Jr. Assessment of fold recognition predictions in CASP6. Proteins 2005;61(Suppl 7):46–66.
5. Read RJ, Chavali G. Assessment of CASP7 predictions in the high accuracy template-based modeling category. Proteins 2007;69(Suppl 8): 27–37.
6. Jauch R, Yeo H, Kolatkar PR, Clarke ND. Assessment of CASP7 structure predictions for template free targets. Proteins 2007;69(Suppl 8): 57–67.
7. Clarke ND, Ezkurdia I, Kopp J, Read RJ, Schwede T, Tress M. Domain definition and target classification for CASP7. Proteins 2007; 69(Suppl 8):10–18.
8. Virnau P, Mirny LA, Kardar M. Intricate knots in proteins: function and evolution. PLoS Comput Biol 2006;2:e122.
9. McDonald IK, Thornton JM. Satisfying hydrogen bonding potential in proteins. J Mol Biol 1994;238:777–793.
10. Zemla A. LGA: a method for finding 3D similarities in protein structures. Nucleic Acids Res 2003;31:3370–3374.

11. Ortiz AR, Strauss CE, Olmea O. MAMMOTH (matching molecular models obtained from theory): an automated method for model comparison. Protein Sci 2002;11:2606–2621.
12. Kryshtafovych A, Fidelis K, Moult J. Progress from CASP6 to CASP7. Proteins 2007;69(Suppl 8):194–207.
13. Venclovas C, Zemla A, Fidelis K, Moult J. Assessment of progress over the CASP experiments. Proteins 2003;53(Suppl 6):585–595.
14. Tress M, Tai CH, Wang G, Ezkurdia I, Lopez G, Valencia A, Lee B, Dunbrack RL, Jr. Domain definition and target classification for CASP6. Proteins 2005;61(Suppl 7):8–18.
15. The CCP4 suite: programs for protein crystallography. Acta Crystallogr D Biol Crystallogr 1994;50(Pt 5):760–763.
16. Wallner B, Elofsson A. Identification of correct regions in protein models using structural, alignment, and consensus information. Protein Sci 2006;15:900–913.
17. Cozzetto D, Kryshtafovych A, Ceriani M, Tramontano A. Assessment of predictions in the model quality assessment category. Proteins 2007;69(Suppl 8):175–183.
18. Bent HA. Uses (and abuses) of models in teaching chemistry. J Chem Educ 1984;61:774–777.

2.3: Predicting the functional impact of mutations

Manuscript in preparation:

Predicting the functional impact of single point mutations: a test of accuracy and robustness of descriptors and machine learning methods.

Abstract

Predicting the effect of single point mutations is of central importance in biology, as it can help rationalise the cause of diseases in a clinical setting and aid in planning mutagenesis in an experimental one. Computational approaches to understanding the effects of mutations have thus gained much interest in recent years. In this study, several classifiers incorporating evolutionary and structural information are devised for the purpose of predicting the effects of single point mutations. Their robustness and training set size dependence is rigorously evaluated using mutation level and protein level cross validation. The top classifiers presented here are validated using systematic mutagenesis data so as to confirm their predictive power. It is found that using evolutionary data, structural features of the mutation sites and mean force potentials to predict stability changes, classifiers can be trained which exceed equivalent currently published ones in performance. We find that mutation level validation overestimates predictive power, whereas protein level validation provides much more robust estimates, which are confirmed by validation on unseen proteins.

Introduction

Predicting the impact of a mutation on the function of a protein is an important issue in biosciences. In the field of medical research, understanding the impact of a mutation on protein structure and function can lend insight into the molecular basis of disease, such as the loss of function of phenylalanine hydroxylase as the cause of phenyl ketonuria[1] or the inactivation of p53 in cancer[2]. The deleterious molecular effects a mutation can have are diverse. They range from abolishing essential residues responsible for catalysis or regulation to altering a protein's folding kinetics or decreasing stability which in turn can cause a severe reduction in the protein's effective concentration or cause aggregation.

Understanding and predicting mutational impact on function is of great importance, therefore there is much interest in developing computational approaches to phenotype prediction. These methods employ information from two primary sources, sequence evolution and molecular structure. Evolutionary information is derived from sequence alignments of homologous proteins. Conservation scores[3] or position specific scoring matrices[4] can be derived from these alignments for the purpose of estimating the probable impact of mutations[5]. Evolutionary information has been used in protein engineering to dramatically increase protein thermostability[6], predicting phenotypic effects of mutations on synthetic datasets[7] and predicting the effect of mutations in humans[8]. Structural information was first used in a large scale manner to understand the impact of human SNPs by Bork and co-workers[9,10], Chasman&Adams[11] and Wang&Moult[12,13], who analysed the structural context of mutations and developed basic classification models to rationalize and predict their effects on the phenotype. In further studies, the biophysical properties of mutation sites were analysed and informative descriptors for predicting mutation phenotypes were identified[14,15,16]. Increasingly complex classification models have been used to classify SNPs using evolutionary and structural data. Initially, rule-based methods using structural descriptors were established for the purpose of predicting the molecular effect of mutations[10,12,17]. Simple rule sets have been superseded by the use of machine learning methods to optimally combine the various descriptors, such as support vector machines[18,19,20,21], decision trees[18], Random Forest based schemes[19,22], and neural nets[23].

However, a number of issues have been raised with regard to their robustness. Early rule-based methods have been shown to be prone to error, making them insufficient for the purpose of medical diagnosis[24]. Others have raised the issue robustness of the classification models, pointing out that training set dependence may limit their general applicability[25]. Here, the relative importance and robustness of evolutionary and structural data to predict mutation phenotypes is assessed. The sequence data comprises sequence profile and conservation scores and in terms of structural information, pseudo-energy functions are used in conjunction with basic biophysical descriptors. Extensive benchmarks are performed to assess information content of the descriptors, and their robustness to training set size variations. In addition, three machine learning methods are compared in terms of their ability to predict mutation phenotypes. The mutation data used to train the classifiers presented here are from Swiss-Prot[26]; these variants are assigned a phenotype based largely on evidence in the scientific literature, but which may in case be confirmed using the biophysical property changes induced by a mutation. While many others have used this set[19,27,28], predicting mutation effects using similar descriptors as were used to annotate their phenotypes may lead to over-estimation of accuracy. To guard against the possibility of circular reasoning, the best classifiers presented here are cross-validated against systematic mutagenesis data to confirm their transferability to new protein classes and to assess the reliability of the accuracy estimates.

Results

Information content of descriptors

Sequence information

Tolerant sites in proteins evolve through a process of neutral evolution, whereas functionally constrained ones show less variation across species. Given a sequence alignment of homologous proteins, two types of information can be extracted by examining the amino acid composition of an alignment, namely the site conservation and the site residue preference. In the present study, conservation was calculated based on a sequence alignment of homologous proteins found by PSI-BLAST, using 2 and 3 iterations respectively; these scores are referred to as the site conservation scores. In addition to site-specific conservation scores, other evolutionary information relevant to SNP phenotype prediction can be derived from the degree of conservation of the entire protein[20], as high conservation of a protein may imply involvement in an important biological function. The average sequence conservation and its standard deviation over all residues in a protein based on both alignments from the two PSI-BLAST searches were calculated and are referred to here as the protein conservation scores. Residue preference was calculated using the program SIFT. The wild-type residue score, the mutant residue score, as well as the difference between the two (referred to here as "delta SIFT"), were used.

To examine the information content of the sequence descriptors, a Random Forest classifier was trained on all mutations in the dataset and their relative contribution was determined using the internal importance estimation functionality of the Random Forest library (for more a definition of the MDA and MDG scores, please see materials and methods). The average importance values of each descriptor are displayed in table 1. The mutant SIFT score and the delta SIFT score ranked highest by a large margin in terms of MDG, but were only of mediocre importance in their MDA score, while the wild type SIFT score was the least informative by both measures. Wild-type and mutant identity scores ranked a distant third and fourth by MDG, and were also mediocre in their performance by MDA. The protein conservation scores ranked highest by MDA, but their MDG scores were mediocre, while the site conservation values were the least informative of the conservation scores, by either measure. Thus there is contrast between the SIFT scores, which scored highest by MDG and the whole protein conservation scores, which scored highest by MDA.

Structure information

Changes in stability can affect protein function and lead to a deleterious phenotype. Knowledge of the effect on stability is likely to be of importance in phenotype prediction and structural information can be used to make estimates about the likelihood of a mutation affecting protein stability. Here, mutations were mapped to protein structures and the properties of the mutation site were used as descriptors in classification. If exact protein structures were not available, homology models from the Swiss-Model Repository[29] were used.

The standard geometric features used for characterising mutation sites included solvent accessibility and the secondary structure type at the position at which the mutation occurs. Here a more advanced method of extracting information from structures was used, namely by making a prediction of the stability change induced by the mutation. Atomic mean force potentials[51], which use the observed frequency of contacts between the different classes of atoms in known protein structures to derive interaction energies, have been shown to be potent tools in assessing changes in stability of mutations[52]. In order to use mean force potentials to estimate stability changes, in-silico mutations have to be performed in order to obtain a structure model for the mutant protein. While tests aimed at determining the most accurate modelling protocol have been performed[53,54], it is not entirely clear which method is the most consistent for the purpose of predicting stability changes. Traditionally, methods are benchmarked by comparing the predicted structure to the experimental one using structural measures such as root mean square deviation (RMSD) or chi angle deviation of the amino acid residues in the vicinity of the mutation site[53,54]. However, it is not clear that structural accuracy is necessarily sufficient in estimating the reliability of predicted energy changes. Here, an alternative criterion is proposed, namely the comparison of modelling protocols based on how informative the resulting mutant structures are in predicting the mutation phenotype. The modelling protocols used are outlined in the materials and methods section of this report and are summarized in table 2.

Information content and ranking of the structural descriptors.

A Random Forest was trained on the complete SNP dataset using the mean force potential pseudo-energies, together with the geometric descriptors (solvent accessibility, secondary structure and the amino acid type of the wild-type and mutant residues). The average importance values of the structural descriptors are reported in table 3. The wild-type and mutant amino acid residue type contained the most information out of the scores used here, as evidenced by their high scores in both MDA and MDG values. It is apparent that the most informative model-building/wild-type combination was the most basic approach, namely using non-energy minimised wild type structures and in-silico mutations performed by Promod-II (protocol “#0”). Both the MDA and MDG averages show that the pseudo-energies from the ANOLEA potential provided the most information of the pseudo-energy methods. The remainder of the MFP protocols had MDA values comparable to those of ANOLEA-#0, their MDG values however were significantly lower. Of all the MFP scores, only the ANOLEA-#0 ranked higher than the solvent accessibility, which was the fourth most important of the descriptors used. It is also worth noting that secondary structure was consistently uninformative, ranking last throughout the training runs.

Robustness of mutant phenotype prediction methods

To test the predictive power of a classifier, it has to be tested on data not used for training. An obvious requirement made for validation is that the data points be independent of one another, which, in the case of many descriptors here, this is not the case. Protein conservation scores, for instance, are the same for all mutations in a given protein. This dependence may overestimate a classifier’s accuracy and not make robust estimates as to how informative the classifier is going to be when applied to new, unseen data. A further point is that the split ratios used for partitioning the data into training and validation sets are arbitrary and vary strongly for classifiers presented in the literature. Some have used ten-fold cross-validation[19], while others have used more rigorous validation such as a 60/40 randomized split[22]. High training set/validation set ratios may give rather optimistic accuracy estimates and not provide good estimates as to the method's performance on new data, nor does it tell the user anything about the robustness of the method to small training set sizes. To validate our classifiers, a randomized split was used to partition the data into training and validation sets. Two schemes were used, mutation level splitting (referred to as RS-M) and protein level splitting (referred to as RS-P). A fixed split ratio of 90/10 was used for the randomized split in order to assess the maximum performance of a classifier and report values for comparison with other published classification models. In addition, training set size dependence, and thus robustness, was investigated by varying the split proportion used between 0 and 1 in small increments. Thus the performance of a classifier can be expressed a function of the split proportion used for creating training and validation sets.

Sequence information

90/10 randomised split. The maximal performance of the sequence classifiers was determined using a 30-fold repetition of a 90/10 randomised split. The results of this validation appear in table 4. For the RS-M scheme, the full sequence classifier, which used both the local descriptors and the whole protein conservation values, (for an outline of the descriptors used here, see table 5) had an MCC far in excess of the local sequence classifier, which used only residue specific data (an MCC of 0.593 vs. 0.494). However, the accuracy dropped to essentially the same level, when performing cross validation by protein. By contrast, the accuracy of the local sequence classifier was practically identical for the RS-M scheme as for the RS-P scheme.

Training set size dependence. In order to further test the learning behaviour of the classifiers, the dependence of the performance on training set size was tested. The result of this validation is displayed in figure 1(A). For the RS-M, the full sequence classifier increased steadily towards its maximum performance with increasing training set size, whereas the local sequence classifier reached a plateau in its performance using approximately 20% of the data. When using RS-P, the local and full sequence classifiers attained the same maximum performance, which is comparable to the performance of the local sequence classifier in the RS-M. It is notable that the full sequence classifier learned more slowly than the local sequence classifier, as is evidenced by the shallower learning curves in figure 1(A).

Structure information

Above, the reliability of the mutant structure models calculated by in-silico mutation protocols were tested for their information content by calculating pseudo-energy differences between them and the wild-types structures. A conservative mutation modelling protocol was shown to be the most informative; nevertheless, the remainder were not uninformative and could still be of value in classification. In order to test the effect of including all pseudo-energies compared to including only the one derived using the most conservative protocol, the two cross-validation schemes were used. Both classifiers used the wild-type and mutant amino acid type, as well as the wild-type residue's relative solvent accessibility and secondary structure type as factors (see table 5, column 1, 2, 12 and 13, respectively). The minimal structure classifier used only the most conservative wild-type/mutant structure combination to derive a single pseudo-energy term (protocol #0, table 2). A second classification model, termed the full structure classifier used the 10 pseudo energy differences derived from the ten mutant/wild-type combinations.

The results of the structure based classifiers for the 90/10 randomized split are displayed in table 4. RS-M validation showed that the full structure classifier attained a level of performance almost comparable to that of the local sequence classifier (a balanced error rate of 27.6% and an MCC of 0.449), albeit at the expense of using 10 partially redundant mean force potential terms. The minimal structure classifier attained a significantly higher error rate of 34.3% and an MCC of 0.314. Using RS-P, the full structure classifier performed only marginally better than the minimal structure classifier, boasting an improvement in the MCC of 0.02. This disparity in the accuracy observed between the two different validation schemes is similar to that observed for the sequence classifiers and implies that a degree of protein specific model fitting is occurring.

Figure 1(B) shows the MCC of the classifiers in relation to the size of the training proportion used. When using the RS-M scheme, the full structure classifier required almost all the set as training data to reach maximum performance, but greatly exceeded the minimal structure classifier in performance. The robustness of the minimal structure classifier is evidenced by the fact it attained near-maximal accuracy using only approximately 20% of the data. When applying RS-P, the two classifiers reached equal maximum performance, which is comparable to that of the minimal structure classifier using RS-M. The full structure classifier not only suffered a great drop in maximal MCC, but the shallow incline of learning curve in figure 1(B) indicates that it required a greater proportion of the data to attain it than the minimal structure classifier does.

Combined sequence and structure information

To determine the level of improvement in the accuracy obtained using structural information in addition to the sequence information, two classifiers combining sequence and structure information were subjected to rigorous analysis. Two models were used (see table 5): the full combined classifier used all terms available whereas the minimal combined classifier combined the terms from the local sequence classifier and the minimal structure classifier and thus represents a minimal combination of purely local terms.

The results of the validation by 90/10 randomised split for the combined models are displayed in table 4. When assessed by RS-M, the maximal performance of the combined full classifier (MCC=0.610) exceeded that of the full sequence classifier (MCC=0.593). The minimal combined classifier, which excludes the protein conservation averages and variances as well as most of the pseudo-energy terms, had a lower performance; the balanced error rate (BER) was reduced by 3.7% and the MCC by almost 7.6%. However, RS-P validation reduced this performance gap between the two to an insignificant level. The learning curves, as shown in figure 1(C), indicate that the full combined classifier had a shallower learning curve than the minimal combined classifier: while it did eventually match the minimal one in performance, it required more data to do so.

To estimate the added value of structural information, the performance and learning behaviour of the full sequence, full structure and the full combined classifiers was compared, as seen in figure 1(D). The combined classifier incorporating full sequence and full structure information exceeded the full sequence classifier by a noticeable margin. The full structure classifier fared comparatively poorly in that it had the shallowest learning curve and the lowest maximum MCC. When using RS-P, the full combined classifier still exceeded that of the full sequence classifier; the performance difference between the two is similar for both mutation level and protein level splitting.

Performance of various machine learning tools

So far, the machine learning method Random Forest has been used for the purpose of classification. The learning behaviour and performance of other methods, the support vector machine and the binary decision tree, were investigated using the two validation schemes presented. All three methods used the descriptors of the full combined classifier, as outlined above. Figure 1(F) shows the learning behaviour of all three methods using both validation schemes. Random Forest ranked highest, using both schemes, followed by the SVM. The disparity in the MCC between the two cross-validation schemes was highest for Random Forest, while the SVM showed a lower, but still significant difference. The decision tree, which performed consistently worse than the other methods, showed little change in performance between the both splitting schemes.

Confirmation of accuracy using unseen data

A further, even more representative test of robustness is validation on a dataset, which has not yet been seen and which represents a wholly new protein class. Here, a classifier trained on a set of human mutations was validated by making predictions on systematic mutation data of unrelated proteins from other organisms. Both lysozyme of the T4 bacteriophage[34] and the Lac repressor protein[35] have been subjected to systematic mutagenesis, whereby most of the sites in the two proteins have been mutated to almost every other residue. These datasets have two benefits. Firstly, a single, objective phenotype is being measured in isolation, thus circumventing the problem of incomplete penetrance of the molecular phenotype in the human SNP dataset. Secondly, all amino acids are tested at all sites of one protein, rather than a small number of mutations in many, very different proteins in the SNP dataset. By examining how well our method can classify the measured effects of mutations in these systems, we can establish how well the molecular phenotype is being learned.

The table 6 shows the results of applying the Random Forest models using sequence and structure data, the full combined classifier and the minimal combined classifier, to the task of predicting the phenotypes observed in the systematic mutagenesis experiments. The highest performance was achieved for the lysozyme mutants at 37°C; the accuracy rates for the full combined classifier were significantly lower than the performance estimates from the RS-M, but only marginally lower than those achieved during the RS-P validation. The minimal combined classifier performs worse than the full combined classifier on both counts. For the lysozyme mutants at 25°C, both classifiers perform worse, the accuracy of the full combined classifier being insignificantly higher than that of the minimal combined classifier. For the lysozyme data the false positive and false negative error rates are more balanced for the full combined classifier than for the minimal combined classifier, although the false positive rate is still considerably lower than the false negative rate. For the Lac Repressor data, the full combined classifier performs marginally worse than the minimal combined classifier, implying that despite the higher performance for the full combined classifier observed during validation on the human SNP set, it doesn't necessarily guarantee the best results for all proteins.

Comparison with other published methods

The comparison with the classification tools of other authors is difficult for three reasons. First, the datasets used for training and validation vary between studies, making them not strictly comparable. Second, the validation schemes differ tremendously, and affect the degree to which performance estimates are comparable. Thirdly, the descriptors differ in their applicability; the use of descriptors drawn from knowledge bases such as Swiss-Prot may improve overall accuracy, but the classifiers do not work when this information is unavailable. Nevertheless, some performance estimates have been published which are derived from mutations from Swiss-Prot. While still not perfectly comparable, they can act as a general guide.

In table 7, the combined classifiers, validated by 90/10 RS-M are compared to other published methods. Bao and Cui[19] use a set of 4013 mutations from Swiss-Prot. Their set is strongly unbalanced with only 532 neutral mutations, the rest being deleterious. This imbalance reflected in their false positive and false negative rates, and is extremely detrimental to their reported BER and MCC. Barenboim and colleagues[22] use a set of 1315 mutation and perform 10-fold cross validation on their statistical geometry based method.

They achieve better error rates than Bao and Cui, owed in part to their more balanced data set. Ye et al [28] use a set of 3438 mutations in 522 proteins and achieve an MCC of 0.604 in five-fold protein level cross-validation. Hu & Yan[8] achieve good results using a small set of sequence based descriptors to train a decision tree using the same data set as Ye et al; their protein level cross-validation gives an MCC of 0.607, but when applying their classifier to the complete Swiss-Prot dataset, their MCC drops to 0.42. In the present study, the mutations from the human variants pages from Swiss-Prot were used to train the classifiers. Using a far more diverse set of proteins, the full combined classifier has an MCC of 0.532; a higher MCC is attainable using mutation level validation (MCC=0.610).

A further comparison with other methods is possible, using the systematic mutation datasets (see table 8). Karchin et al[21] used a support vector machine based on structural and evolutionary data to classify mutations in the Lac repressor and T4 lysozyme mutagenesis datasets. By training on one dataset and validating on the other, they obtained reliable estimates of their classifiers' accuracy. Their on average best method obtained an accuracy of 65.8% (error rate=34.2%) on Lac repressor and 71.4% (error rate=28.6%) for lysozyme; it was outperformed by the full combined classifier presented here. It is worth noting that Karchin et al. were able train better predictors for the datasets individually; they could attaining an accuracy of 66.8%, on the Lac repressor data (corresponding to an error rate of 33.2%) using a minimal set of only three descriptors, and an accuracy of 74.8% (corresponding to an error rate of 25.2%) on the lysozyme data, using a set of 32 descriptors. Both these methods performed well only on one dataset and were less accurate on average than their top predictor was. Bromberg and Rost[23] also validated their method using the lysozyme and Lac repressor data and reported their overall accuracy (defined as the number of correct predictions divided by the sample size, which corresponds to 1 minus the error rate reported here). For the Lac repressor and lysozyme mutagenesis data, they report an accuracy of 72.7% (error rate = 27.3%) and 73.2 (error rate = 26.8%). Their basic model has a reported accuracy of 70.7 (ER=29.3) and 70.0 (ER=30.0) for the Lac repressor and lysozyme datasets, respectively. The minimal combined classifier compares favourably with their non-annotated version, outperforming it on T4 lysozyme, but performing less accurately on Lac Repressor. The same is observed in the comparison of the full combined classifier with their annotated model: it performs better on lysozyme, but worse on Lac repressor.

The issue of descriptors assigned by protein

As mentioned above, some of the descriptors were assigned on a per-protein basis, i.e. for each mutation occurring in a given protein the same value was assigned; this is the case for the whole protein conservation scores. However, machine learning methods generally assume the independence of the values used for prediction. If this is not given, there is scope for over-training on the dataset, particularly when the validation techniques are not rigorous. The inclusion of per-protein factors leads to a dramatic increase in the apparent predictive power of the method, but it is unclear whether or not the increase in performance is due to a biologically explainable effect or whether it is merely due to non-stringent validation techniques. Here, random factors were used in order to test whether a similar level of predictive power could be achieved.

Random factors were assigned on a protein-wise basis, in the same manner as were the mean and standard deviation of the protein conservation. For each protein, four random real numbers between 0 and 1 were generated and assigned to each mutant from this protein. The effects of replacing the whole protein conservation scores are shown in figure 1(E). Again, the learning effects were examined using RS-M and RS-P schemes for dataset partitioning. The local sequence classifier was included for reference, as a base line of performance. Using RS-M, the random effects showed a marked improvement over the local classifier, although the classifier incorporating them didn't reach the accuracy of the protein conservation score based method (full sequence classifier). The random effects displayed the same behaviour when switching from RS-M to RS-P, as did the whole protein conservation scores: they suffered a performance drop to level of that of the local sequence classifier and their learning curves became shallower.

Discussion

Sequence based descriptors. Sequence information is highly informative for the purpose of scoring SNPs. The Random Forest importance scores show that the mutant SIFT value and delta SIFT value, ranked highest according to the MDG value and do so by a considerable margin. The protein conservation scores have the highest MDA values, but only by a small margin, but their MDG values are only mediocre. The site conservation values are surprisingly uninformative compared to the other descriptors, scoring badly in terms of both MDG and MDA. The two different splitting schemes, one by leaving out a given proportion of the mutations and the other by leaving out all mutations from a given proportion of the proteins, are employed to simulate a realistic situation, where the protein in which a mutation of interest occurs, has not yet been seen. When employing RS-M, the full sequence classifier significantly outperformed the local sequence classifier, whereas with RS-P, the two classifiers performed equally well, with their performance being comparable to that of the local sequence classifier as assessed by RS-M. This shows that the local sequence classifier does not need protein specific data to reach its maximum performance, whereas the full classifier does. The information learned from the average conservation of the protein provides little benefit when applied to other proteins. In addition, for RS-P, the full sequence classifier required more data to attain the same performance as the local method.

Structure modelling protocols and mean force potentials. Structural information differs in its information content from evolutionary data, in that it provides mainly information pertaining to stability, rather than providing specific information on biological function. The use of structural data is investigated as is the use of mean force potentials to calculate the energy difference between the wild-type and mutant protein. The importance scores output by Random Forest show that the wild-type/mutant structure from mutation protocol #0 performed best. It appears that elaborate modelling on average only decreases the utility of the structure models, and that the most parsimonious approach, i.e. changing the structures as little as possible, is the most effective. Only the pseudo-energy differences derived using the simplest modelling protocol ("#0") were more informative than the solvent accessibility values and thus contained information beyond a residue's propensity to burial; nevertheless the use of pseudo-energy terms from other mutation protocols was beneficial to the overall predictive power of the classifier. Secondary structure was little informative on the mutation dataset, despite the fact it contains implicit information about the backbone geometry of the mutation site.

The issue of robustness of the structure based classifiers was assessed using the same two splitting schemes, RS-M and RS-P. The classifier incorporating all 10 pseudo-energy terms was more effective than that using only 1 pseudo-energy, when using RS-M. Using a 90/10 randomised split showed that the average MCC was significantly higher for the full structure classifier compared to that of the minimal structure classifier; its performance was close to that of the local sequence classifier. However, RS-P indicates that using 10 pseudo-energies provides little benefit over using a single pseudo-energy term when applied to unseen proteins. The learning curve further shows that the additional pseudo-energies decrease the rate of learning when using RS-P. The effect is remarkably similar to performance of the full sequence classifier, which uses protein conservation descriptors, compared to the local sequence classifier: the full sequence classifier attained a higher maximum RS-M performance, was but its rate of learning for RS-P was slower.

Combined sequence and structure classifiers. The added value of structure data over sequence data is assessed by combining both sets of terms into a single classifier. Using RS-M, the full combined classifier showed a notable increase in accuracy over the full sequence classifier; it didn't however show the same degree of training set size dependence as the full structure classifier did. RS-P shows the same steady increase in performance of the full combined classifier over the full sequence classifier, showing that neither is more robust than the other.

Machine learning tools. Three machine learning methods have been benchmarked in the present work, Random Forest, SVM and decision tree. Random Forest outperformed both the other machine learning methods by a notable margin. The performance of the SVM was lower than that of the Random Forest, for both splitting schemes; the decision tree generally performed poorly compared to other methods, as presumably its conceptual simplicity did not allow it to be fit to complex patterns in the data. This simplicity however made decision trees very robust, as is evident from the similarity in its performance between RS-M and RS-P. It also required only little data to reach its maximum performance during RS-M, even if this was significantly lower than for the other two methods.

Confirming the performance and performance estimates on unseen data. To further validate our Random Forest classifiers and to test the robustness of the accuracy estimates, the full and the minimal combined classifier were used to classify mutation data from systematic mutation datasets. The accuracy of the classifiers on these datasets was generally lower than that obtained during self-consistency validation on the Swiss-Prot variant set. The most accurately classified set was the lysozyme-37 data; the full combined classifier's accuracy approaches the accuracy achieved during the self-validation using RS-P. This supports the notion that RS-P provides more realistic estimates of the expected accuracy than RS-M. The minimal combined classifier was less accurate, which would initially indicate a gain in predictive power due to the additional descriptors included in the full combined classifier.

For both classifiers, the classification performance for the lysozyme-25 data was somewhat lower than the accuracies according to RS-M and RS-M, whereby again the full combined classifier outperformed the minimal one, albeit by a negligible margin. The reduction in performance of both classifiers was presumably due to the influence of temperature on those mutations which cause a reduction in stability. At lower temperature the destabilisation of a protein will be less prominent and its stability will be sufficient for the effect of the mutation to go unnoticed. Due to this masking effect, a correctly predicted loss of stability will be seen as a classification error and lead to an increase in the false positive rate.

The results for Lac repressor dataset displayed highly imbalanced false positive and false negative rates. This protein binds DNA as a tetramer, but the structural data is only taken from the monomeric form of the structure. As solvent accessibility implies tolerance to mutation, residues at the binding interfaces will have misleading solvent accessibility values. Here the minimal combined classifier fared marginally better than the full combined classifier, which may be indicative of increased robustness of the classifier to the limitations in the structural data. The general observation that none of the validation performance estimates approached those observed during self-validation using the RS-M scheme implies that it is a less robust estimator of the true performance than the RS-P scheme. Even more accurate estimates could presumably be achieved using even more rigorous validation schemes. For instance, validation at a protein family level, rather than just at the protein level could remove bias due to family specific traits being learned during training.

Comparison with other published methods. A minimalist approach has been presented by Bao and Cui [19], who use structural and evolutionary information to predict phenotypes. Our full combined classifier exceeded their performance, both when using mutation level and protein level validation. An unfortunate drawback to their study was the imbalance of deleterious and neutral SNPs, which affected their reported MCC values. By contrast, Dobson et al[27] used a fully balanced set for training their prediction tool and reported an MCC of 0.49 for their set of 3821 SNPs. The application of their classifier is limited by the requirement of Swiss-Prot annotation, which is not universally available. By contrast, our full combined classifier is independent of annotation, yet still achieved an MCC of 0.61 and 0.532 using mutation level and protein level cross-validation, respectively.

Others (Ye at al[28], Hu and Yan[8]) achieved higher MCC values than our method using protein level cross validation. The robustness of this estimate is however unclear, as Hu and Yan's classifier achieved an MCC of only 0.42 when validated on a larger, unseen set of Swiss-Prot mutations (21185 mutations, in their case). Both studies use a binary descriptor for classification, which indicates whether or not the protein a mutation occurs in is a member of the human leukocyte antigen (HLA) family. Ye et al argued that it allows specific properties of this family to be learned; however, performance estimates for only the HLA proteins would have to be presented to confirm this assertion. Our classifiers do not have such drawbacks as they use no categorical family specific information. The minimal combined classifier learned fast, needing only a very small proportion of the data to reach maximal accuracy and thus circumvents vastly over-optimistic accuracy rates. Furthermore, our method achieved higher performance on the lysozyme mutagenesis data than Hu&Yan do on Swiss-Prot data, even though lysozyme constitutes an unseen protein from a different organism.

Validation on the systematic mutagenesis data from Lac repressor and T4 lysozyme presents allows an objective comparison to other methods. Karchin et al[21] trained an SVM using evolutionary and structural data on either Lac repressor or T4 lysozyme, and validated their classifier on the other dataset. Here, the full combined classifier was shown to outperform their on average best classifier. Karchin et al were able to reach higher accuracy on the individual datasets using different descriptor combinations for their classifiers, but the improvement came at the expense of a loss of accuracy when the classifier was validated on the other set. The accuracy of Bromberg and Rost's method[23] depended on the inclusion of specific annotation information; using annotation their method outperformed our minimal combined classifier, when omitting annotation their method performed worse on lysozyme, but still better on Lac repressor. On lysozyme dataset, the full combined classifier outperformed both the annotated and non-annotated classifiers of Bromberg and Rost, but was inferior on the Lac repressor set. This imbalanced behaviour is corresponds to the behaviour observed by Karchin et al, who obtained high accuracy on one dataset, by sacrificing it on the other; balancing the performance between the datasets appears to entail a loss in predictive power.

The effect of random descriptors assigned by protein. The performance of the classifier using protein level random factors showed similar behaviour to that of the full sequence classifier, which uses protein conservation scores: it showed a large improvement over the local sequence classifier, which uses only local sequence information. This improvement was, however, not as great for the random values as for the protein conservation scores. This shows that a large proportion of the accuracy improvement by using per protein descriptors is due to family specific fitting; the values can be arbitrary and yet still allow Random Forest to derive information from them. The fact that the 4 whole protein conservation scores still performed better than the random effects (using 90/10 RS-M, the classifier using 4 conservation scores attained a MCC of 0.586, where as the classifier using 4 random effects achieved an MCC of 0.556) shows that they still do contain information which is useful in classification. Like the profile methods, the random effects attained similar performance as the local sequence classifier using RS-P, but the learning curve was shallower. In all, the protein conservation averages and standard deviations contain more information than the random descriptors, but not as much as one would assume when comparing their performance to that of the local sequence classifier based on the RS-M accuracy estimates. The inclusion of data on protein conservation averages is merited if the training dataset is large enough for the slower rate of learning not to be felt.

Summary

In this work, methods for predicting the effect of mutations on the phenotype are rigorously validated. The use of structural information improves the overall performance compared to using only evolutionary data, confirming the utility of protein structure in mutation effect prediction. While the top classifier presented here, which uses sequence and structure information, exceeds other comparable methods validated on SNP datasets, the performance depends crucially on the type of validation used. Validation techniques which do not account for statistical couplings in the data may overestimate the performance. Protein level validation shows that a method using the minimal subset of descriptors can achieve a similar level of performance as a method using the full set of descriptors.

Materials and methods

Swiss-Prot human variation dataset. The human variants listings of the Swiss-Prot database [26] were used, as supplied on the Uniprot web site (http://www.uniprot.org/docs/humsavar). In the version used (UniProt Release 55.0) there were 44208 distinct mutations in 8658 different proteins. For 38% of these, i.e. 17016 mutations in 3791 proteins, structural information was available either in form of experimental structures or homology models from the Swiss-Model Repository[29]. Of the mapped mutations, 58.1% (9891 mutations) were disease causing and the remainder (7125 mutations) were annotated as polymorphisms. Mutations without a confirmed phenotype ("Unclassified") were excluded from the present analysis.

Systematic mutagenesis datasets. Two systematic mutagenesis datasets were used for validation purposes. The complete mutagenesis of bacteriophage T4 lysozyme[34] at two different temperatures (25°C and 37°C) provided 4030 mutations (2015 mutations at each temperature). For the lysozyme data, mutations were classed as neutral only if full wild-type activity was not reported, otherwise they were considered deleterious. The Lac repressor protein mutagenesis dataset [35] provided 4000 mutations. Here, only mutations which did not have full wild-type activity were deemed to be deleterious, otherwise they are classed as neutral.

Sequence descriptors. For calculating conservation scores, sequence data was collected using a PSI-BLAST[32] (version 2.2.16) search against the Uniref90 database[33]. For each sequence, 2 searches were performed using 2 and 3 iterations (PSI-BLAST parameter "j"). The search employed an iterative inclusion e-value cut off (PSI-BLAST parameter "h") of 10^{-5} and a final cut-off (PSI-BLAST parameter "e") of 10^{-3}. PSI-BLAST hits with less than 80% coverage of the target sequence were removed from the alignment. Conservation scores were calculated based on the BLAST alignment and using the program Scorecons[3] with the default ("valdar01") weighting scheme. SIFT[7] was run using the Uniref90 database. The SIFT scores of the wild-type residue, that of the mutant residue and the difference between the two were calculated. In addition to the evolutionary data, the identity of the wild-type and the mutant residue are used as categorical descriptors during classification.

Structural coverage and homology models. For SNP analysis, structural data were obtained from structures in the PDB[34]. In cases where no direct structure was obtainable, homology models were taken from the Swiss-Model Repository[29] a database of structures built using the Swiss-Model pipeline[35]. As multiple fragments may be available for each sequence, the fragment with the highest target template identity (over the length of the covered region) was used.

Structural information. Using structures or homology models, the solvent accessibility was calculated using the program NACCESS[36], whereby the relative solvent accessibility was used (i.e. the solvent accessibility relative to that of the same residue in a Gly-X-Gly trimer, where X is the amino acid). Secondary structure was assigned using DSSP[37].

In-silico mutation and energy difference calculation. 5 protocols were used to generate structure for the mutant protein. The first employs Promod-II[38] to mutate the site of interest, followed by an energy minimisation of only the mutated residue using the GROMOS force field[39]. The second uses simple side chain replacement by SCWRL (version 3.0[40]), the third additionally includes 50 rounds of energy minimisation, using the GROMOS force field implemented in Promod-II, of the whole structure following mutation. The fourth and fifth protocol use SCWRL to mutate the appropriate residue and to repack the side chains which have at least one atom within 10 Angstroms of the mutated residue's side chain, whereby protocol 4 does not perform and energy minimisation and protocol 5 subjects the whole protein structure model to 50 rounds of energy minimisation using the GROMOS force field. The five mutation protocols and 2 wild type structures yield 10 combinations of wild-type and mutant pairs, which were analysed using 2 mean force potentials, resulting in 20 combinations. Two wild type structures were used in the calculation of the pseudo-energies, of which one was energy minimised using the GROMOS implementation in Promod-II, while the other was not. The ANOLEA mean force potential[41] was used to calculate pseudo-energies. It uses 40 atom types to represent the 167 atoms from all standard residues in proteins and incorporates a knowledge based solvation term. The difference in pseudo-energy between the wild-type and the mutant structure was used as a descriptor for classification (columns 14-23 in table 5).

Descriptor importance. The relative information content of the descriptors used for mutation classification was assessed by using the importance values provided by the Random Forest classifiers trained on the entire Swiss-Prot SNP dataset. Random Forest performs classification using an ensemble of classification trees, each built on a subset of the data. It provides two measures as to how informative the descriptors used are: the "Mean Decrease Gini" value (MDG) is a measure of a variable's average contribution to class separation at the individual nodes of the trees. The difference in node impurity, as defined by the Gini index (ref), is calculated before and after the split on a particular variable, and is averaged over all nodes at which this variable is used. The "Mean Decrease Accuracy" (MDA) reflects the average difference in accuracy of a tree before and after the values of the variable being analysed are permutated. The accuracy determined on the subset of data used to train the tree and the MDA is the average over all trees (for more details the reader is referred to the Random Forest manual[49]). Random Forest relies on a large number of decision trees which are built using a degree of randomness and both its performance and the importance of the variables vary between different classifiers trained on the same dataset. Consequently, a number of repetitions of the training process need to be performed in order to get a reliable estimate of the importance of the factors used in classification. Here, the process is repeated 30 times and the importance values of each variable are recorded.

Classification accuracy. Based on the number of true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN), various performance scores can be calculated. A classifier's error rate is defined as the number of incorrectly classified data points divided by the total sample size:

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

For unbalanced datasets, i.e. those in which one class is more frequent than the other, two more robust scores of accuracy are available.

The balanced error rate (abbreviated to BER) is defined as:

$$BER = 0.5 \times \left(\frac{FN}{FN + TP} + \frac{FP}{FP + TN} \right)$$

Matthew's correlation coefficient [50] (abbreviated to MCC) is defined as:

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

Two additional measures used here are the false positive rate (FPR) and the false negative rate (FNR):

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

Machine learning methods. All statistical modelling was performed using the R programming language[42]. Implementations of the various machine learning methods are available for R: Random Forest[43] is implemented in the randomForest library[44], support vector machines[45] are supplied as part of the package e1071[46] and binary decision trees[47] are provided in the "tree" package[48].

Validation. Classifier validation was performed using random splitting of the data. For mutation level validation, a specified proportion of the data were drawn without replacement from the sample, and used for training the classifier. The remainder were used for validation. The process was the same for protein level validation, only random drawing a proportion from the set of unique proteins and pooling all mutations from proteins in the sample. This process was repeated 30 times and the mean performance values were recorded. For the systematic mutagenesis datasets, the 30 Random Forests were trained using the entire Swiss-Prot human variation dataset and used to predict the effect of mutations on the mutagenesis data; the average performance values were reported.

References

1. Williams, R.A., Mamotte, C.D. & Burnett, J.R. Phenylketonuria: an inborn error of phenylalanine metabolism. The Clinical Biochemist. Reviews / Australian Association of Clinical Biochemists 29, 31-41 (2008).
2. Sidransky, D. & Hollstein, M. Clinical implications of the p53 gene. Annual Review of Medicine 47, 285-301 (1996).
3. Valdar, W.S.J. Scoring residue conservation. Proteins 48, 227-41 (2002).
4. Ng, P.C. & Henikoff, S. Predicting deleterious amino acid substitutions. Genome research 11, 863-74 (2001).
5. Ng, P.C. & Henikoff, S. Accounting for human polymorphisms predicted to affect protein function. Genome Research 12, 436-46 (2002).
6. Lehmann, M. & Wyss, M. Engineering proteins for thermostability: the use of sequence alignments versus rational design and directed evolution. Current Opinion in Biotechnology 12, 371-5 (2001).
7. Ng, P.C. & Henikoff, S. SIFT: Predicting amino acid changes that affect protein function. Nucleic acids research 31, 3812-4 (2003).
8. Hu, J. & Yan, C. Identification of deleterious non-synonymous single nucleotide polymorphisms using sequence-derived information. BMC Bioinformatics 9, 297 (2008).
9. Sunyaev, S., Ramensky, V. & Bork, P. Towards a structural basis of human non-synonymous single nucleotide polymorphisms. Trends in genetics : TIG 16, 198-200 (2000).
10. Sunyaev, S. et al. Prediction of deleterious human alleles. Human molecular genetics 10, 591-7 (2001).
11. Chasman, D. & Adams, R.M. Predicting the functional consequences of non-synonymous single nucleotide polymorphisms: structure-based assessment of amino acid variation. Journal of molecular biology 307, 683-706 (2001).
12. Wang, Z. & Moult, J. SNPs, protein structure, and disease. Human mutation 17, 263-70 (2001).
13. Wang, Z. & Moult, J. Three-dimensional structural location and molecular functional effects of missense SNPs in the T cell receptor Vbeta domain. Proteins 53, 748-57 (2003).
14. Terp, B.N. et al. Assessing the relative importance of the biophysical properties of amino acid substitutions associated with human genetic disease. Human mutation 20, 98-109 (2002).
15. Ferrer-Costa, C., Orozco, M. & de la Cruz, X. Characterization of disease-associated single amino acid polymorphisms in terms of sequence and structure properties. Journal of molecular biology 315, 771-86 (2002).
16. Stitziel, N.O. et al. Structural location of disease-associated single-nucleotide polymorphisms. Journal of molecular biology 327, 1021-30 (2003).
17. Saunders, C.T. & Baker, D. Evaluation of structural and evolutionary contributions to deleterious mutation prediction. Journal of molecular biology 322, 891-901 (2002).
18. Krishnan, V.G. & Westhead, D.R. A comparative study of machine-learning methods to predict the effects of single nucleotide polymorphisms on protein function. Bioinformatics (Oxford, England) 19, 2199-209 (2003).
19. Bao, L. & Cui, Y. Prediction of the phenotypic effects of non-synonymous single nucleotide polymorphisms using structural and evolutionary information. Bioinformatics (Oxford, England) 21, 2185-90 (2005).
20. Yue, P. & Moult, J. Identification and analysis of deleterious human SNPs. Journal of molecular biology 356, 1263-74 (2006).
21. Karchin, R., Kelly, L. & Sali, A. Improving functional annotation of non-synonomous SNPs with information theory. Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing 397-408doi:15759645 (2005).
22. Barenboim, M. et al. Statistical geometry based prediction of nonsynonymous SNP functional effects using random forest and neuro-fuzzy classifiers. Proteins (2008).

23. Bromberg, Y. & Rost, B. SNAP: predict effect of non-synonymous polymorphisms on function. Nucleic acids research 35, 3823-35 (2007).
24. Tchernitchko, D., Goossens, M. & Wajcman, H. In silico prediction of the deleterious effect of a mutation: proceed with caution in clinical genetics. Clinical Chemistry 50, 1974-8 (2004).
25. Care, M.A. et al. Deleterious SNP prediction: be mindful of your training data! Bioinformatics (Oxford, England) 23, 664-72 (2007).
26. Yip, Y.L. et al. Retrieving mutation-specific information for human proteins in UniProt/Swiss-Prot Knowledgebase. Journal of Bioinformatics and Computational Biology 5, 1215-31 (2007).
27. Dobson, R.J. et al. Predicting deleterious nsSNPs: an analysis of sequence and structural attributes. BMC Bioinformatics 7, 217 (2006).
28. Ye, Z. et al. Finding new structural and sequence attributes to predict possible disease association of single amino acid polymorphism (SAP). Bioinformatics (Oxford, England) 23, 1444-50 (2007).
29. Kiefer, F. et al. The SWISS-MODEL Repository and associated resources. Nucleic Acids Research doi:gkn750 (2008).
30. Rennell, D. et al. Systematic mutation of bacteriophage T4 lysozyme. Journal of molecular biology 222, 67-88 (1991).
31. Suckow, J. et al. Genetic studies of the Lac repressor. XV: 4000 single amino acid substitutions and analysis of the resulting phenotypes on the basis of the protein structure. Journal of molecular biology 261, 509-23 (1996).
32. Altschul, S.F. et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research 25, 3389-402 (1997).
33. Suzek, B.E. et al. UniRef: comprehensive and non-redundant UniProt reference clusters. Bioinformatics (Oxford, England) 23, 1282-8 (2007).
34. Berman, H.M. et al. The Protein Data Bank. Acta crystallographica. Section D, Biological crystallography 58, 899-907 (2002).
35. Schwede, T. et al. SWISS-MODEL: An automated protein homology-modeling server. Nucleic acids research 31, 3381-5 (2003).
36. Hubbard, S.J. & Thornton, J.M. NACCESS.
37. Kabsch, W. & Sander, C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22, 2577-637 (1983).
38. Guex, N. & Peitsch, M.C. SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. Electrophoresis 18, 2714-23 (1997).
39. van Gunsteren, W.F. & Berendsen, H.J.C. Groningen Molecular Simulation (GROMOS) Library Manual. Biomos, Groningen (1987).
40. Canutescu, A.A., Shelenkov, A.A. & Dunbrack, R.L. A graph-theory algorithm for rapid protein side-chain prediction. Protein Science: A Publication of the Protein Society 12, 2001-14 (2003).
41. Melo, F. & Feytmans, E. Assessing protein structures with a non-local atomic interaction energy. Journal of molecular biology 277, 1141-52 (1998).
42. R Development Core Team R: A Language and Environment for Statistical Computing. at <http://www.R-project.org>
43. Breiman, L. Random Forests. Machine Learning 45, 5-32 (2001).
44. Classification and Regression by randomForest . R News 2, 18-22 (2002).
45. Vapnik, V. Statistical Learning theory. (1998).
46. Dimitriadou, E. et al. e1071: Misc Functions of the Department of Statistics.
47. Breiman, L. et al. CART: Classification and Regression Trees. Wadsworth: Belmont, CA (1983).
48. Ripley, B. tree: Classification and regression trees.
49. Breiman, L. Using random forests. At <ftp://ftp.stat.berkeley.edu/pub/users/breiman/Using_random_forests_v4.0.pdf>

50. Matthews, B.W. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochimica Et Biophysica Acta 405, 442-51 (1975).

51. Sippl, M.J. Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins. Journal of Molecular Biology 213, 859-83 (1990).

52. Zhou, H. & Zhou, Y. Stability scale and atomic solvation parameters extracted from 1023 mutation experiments. Proteins 49, 483-92 (2002).

53. Bordner, A.J. & Abagyan, R.A. Large-scale prediction of protein geometry and stability changes for arbitrary single point mutations. Proteins 57, 400-13 (2004).

54. Feyfant, E., Sali, A. & Fiser, A. Modeling mutations in protein structures. Protein Science: A Publication of the Protein Society 16, 2030-41 (2007).

Tables

Factor	Mean Decrease Accuracy	Mean Decrease Gini
Mutant residue SIFT score	0.2697141	1340.8982
Delta SIFT score	0.265745	1190.1542
Mutant residue identity	0.2652595	800.3748
Wild type residue identity	0.2671678	721.3913
Protein conservation Standard Deviation (3 iterations)	0.2703881	671.5299
Site conservation (2 iterations)	0.2581044	650.8633
Protein conservation Standard Deviation (2 iterations)	0.2716731	642.8701
Mean protein conservation (3 iterations)	0.2723504	637.7788
Mean protein conservation (2 iterations)	0.2716628	628.4365
Site conservation (3 iterations)	0.2581864	546.602
Wild type residue SIFT score	0.2234823	431.0573

Table 1: Importance of the factors used in classification, derived by training Random Forest, using all sequence descriptors, on all data points in the Swiss-Prot human variation sample. The average values recorded during 30 repetitions of this procedure are displayed and ordered by their Mean Decrease Gini values.

Protocol	Wild type global energy minimisation	Mutation Protocol	Mutant global energy minimisation
#0	-	Promod-II	-
#1		SWRL	-
#2			+
#3		SCWRL with local repacking	-
#4			+
#5	+	Promod-II	-
#6		SWRL	-
#7			+
#8		SCWRL with local repacking	-
#9			+

Table 2: Summary of the protocols used for calculating wild-type – mutant structure pairs, as used for energy difference calculation.

Factor	Mean Decrease Accuracy	Mean Decrease Gini
Mutant residue identity	0.2675565	887.5957
Wild type residue identity	0.2721382	810.9856
Anolea-0	0.2675859	805.7428
Relative solvent accessibility	0.2722856	763.7696
Anolea-9	0.2514838	565.0958
Anolea-1	0.2662802	555.8894
Anolea-3	0.2475920	542.4026
Anolea-2	0.2561249	531.8517
Anolea-6	0.2649048	525.7636
Anolea-7	0.2586813	522.5529
Anolea-5	0.2625828	490.3038
Anolea-4	0.2587994	474.9356
Anolea-8	0.2601427	465.4379
Secondary structure	0.2226180	340.6063

Table 3: The importance of the various structure features and pseudo-energies in classifying mutation phenotypes in the Swiss-Prot human variation set using Random Forest are displayed. The values are derived by averaging the importance values over 30 repetitions of the training procedure using the entire dataset.

Information	Model	Randomised Split Scheme (by Mutant/Protein)	FPR	FNR	ER	BER	MCC
Sequence	Local	M	0.215	0.290	0.246	0.253	0.494
		P	0.222	0.294	0.253	0.258	0.483
	Full	M	0.173	0.234	0.198	0.203	0.593
		P	0.253	0.238	0.249	0.246	0.502
Structure	Minimal	M	0.284	0.402	0.333	0.343	0.314
		P	0.292	0.396	0.336	0.344	0.311
	Full	M	0.231	0.320	0.268	0.276	0.449
		P	0.338	0.328	0.333	0.333	0.331
Combined	Minimal	M	0.195	0.271	0.227	0.233	0.534
		P	0.213	0.268	0.236	0.240	0.516
	Full	M	0.159	0.232	0.190	0.196	0.610
		P	0.226	0.237	0.230	0.232	0.532

Table 4: The result of cross validation using a 90/10 split for creating training/test sets from the data. The two cross validation schemes involve forming subsets either by sampling at the mutation level or at the protein level.

	Amino acid		Sequence									Structure											
	Residue type		Residue preference(SIFT)			Conservation (2 iterations)			Conservation (3 iterations)			General		ANOLEA Mutation/wild type combination									
Model	Wt	Mutant	Wt	Mutant	Delta	Residue	Mean	Sigma	Site	Mean	Sigma	SASA	DSSP	0	1	2	3	4	5	6	7	8	9
Local Sequence	X	X	X	X	X	X			X														
Full Sequence	X	X	X	X	X	X	X	X	X	X	X												
Minimal Structure	X	X										X	X	X									
Full Structure	X	X										X	X	X	X	X	X	X	X	X	X	X	X
Minimal Combined	X	X	X	X	X	X			X			X	X	X									
Full Combined	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X

Table 5: The descriptors used for classification in the various models. The wild-type ("Wt") and mutant residues (columns 1, 2) are included in all models. The residues preference scores are calculated by SIFT (columns 3-5) for the wild-type ("Wt") and the mutant, the difference is referred to as "Delta". The site conservation, whole protein conservation mean and standard deviation ("Sigma") over all residues in the whole protein are listed. Conservation scores are calculated using multiple sequence alignments from PSI-BLAST, generated using 2 and 3 iterations (columns 6-8 & 9-11 respectively). The acronyms SASA and DSSP (columns 12 and 13) denote solvent accessible surface area and secondary structure, respectively. The ANOLEA pseudo-energy scores (columns 14-23) are calculated on wild-type/mutant structure pairs calculated as outlined in table 2.

Model	Validation	FPR	FNR	ER	BER	MCC
Combined full	90/10 RS-M	0.159	0.232	0.190	0.196	0.610
	90/10 RS-P	0.226	0.237	0.230	0.232	0.532
	Lysozyme (37°C)	0.224	0.285	0.266	0.255	0.461
	Lysozyme (25°C)	0.280	0.356	0.313	0.303	0.361
	Lac Repressor	0.190	0.395	0.327	0.293	0.391
Combined minimal	90/10 RS-M	0.195	0.271	0.227	0.233	0.534
	90/10 RS-P	0.213	0.268	0.236	0.240	0.516
	Lysozyme (37°C)	0.200	0.338	0.294	0.269	0.432
	Lysozyme (25°C)	0.241	0.372	0.334	0.307	0.353
	Lac Repressor	0.217	0.346	0.303	0.281	0.411

Table 6: Confirming of the performance estimates and predicted error rates by means of validating on unseen data. For reference, the results from the self-consistency 90/10 random split validation on the Swiss-Prot variants have been included (italic).

Method	Validation method	Requires annotation	Sample size	FPR	FNR	ER	BER	MCC
Bao & Cui[19]	10-fold CV	No	4013	0.378	0.206		0.292	0.315
Bao & Cui[19]	Unseen data (Swiss-Prot)	No	205	0.300	0.240		0.270	0.352
Barenboim et al.[22]	10-fold CV	No	1919				0.294	0.436
Dobson et al. [27]	10-fold CV	Yes	1218					0.490
Ye et al.[28]	5-fold protein level CV	Yes	3438					0.604
Hu & Yan[8]	4-fold protein level CV	Yes	3438					0.607
Hu & Yan[8]	Unseen data (Swiss-Prot)	Yes	21185					0.420
Combined full model	protein level 90/10 RS	No	17016	0.226	0.237	0.230	0.232	0.532
Combined full model	mutation level 90/10 RS	No	17016	0.159	0.232	0.190	0.196	0.610

Table 7: Comparison to other published methods applied to human variants from Swiss-Prot. Cross validation is abbreviated to CV, randomised split to RS. Cross-validations reported by others were performed at the mutation level, unless explicitly stated otherwise.

Method	Lysozyme	Lac Repressor
Ng & Henikoff[4] - BLOSUM62	47.0%	54.2%
Ng & Henikoff[4] - SIFT	63.3%	68.3%
Karchin et al. [21] – “Top five features”	71.4%	65.8%
Bromberg & Rost[23] – SNAP without annotation	70.0%	70.7%
Bromberg & Rost[23] – SNAP with annotation	73.2%	72.7%
Combined minimal	70.6%	69.7%
Combined full	73.4%	67.3%

Table 8: Comparison of the accuracy of the combined classifiers to other published methods on T4 lysozyme and Lac repressor systematic mutagenesis data. The accuracy is defined as 1 minus the error rate. The values for the other methods were taken from the respective publications, the reader is referred to these for details of the methods.

Figures

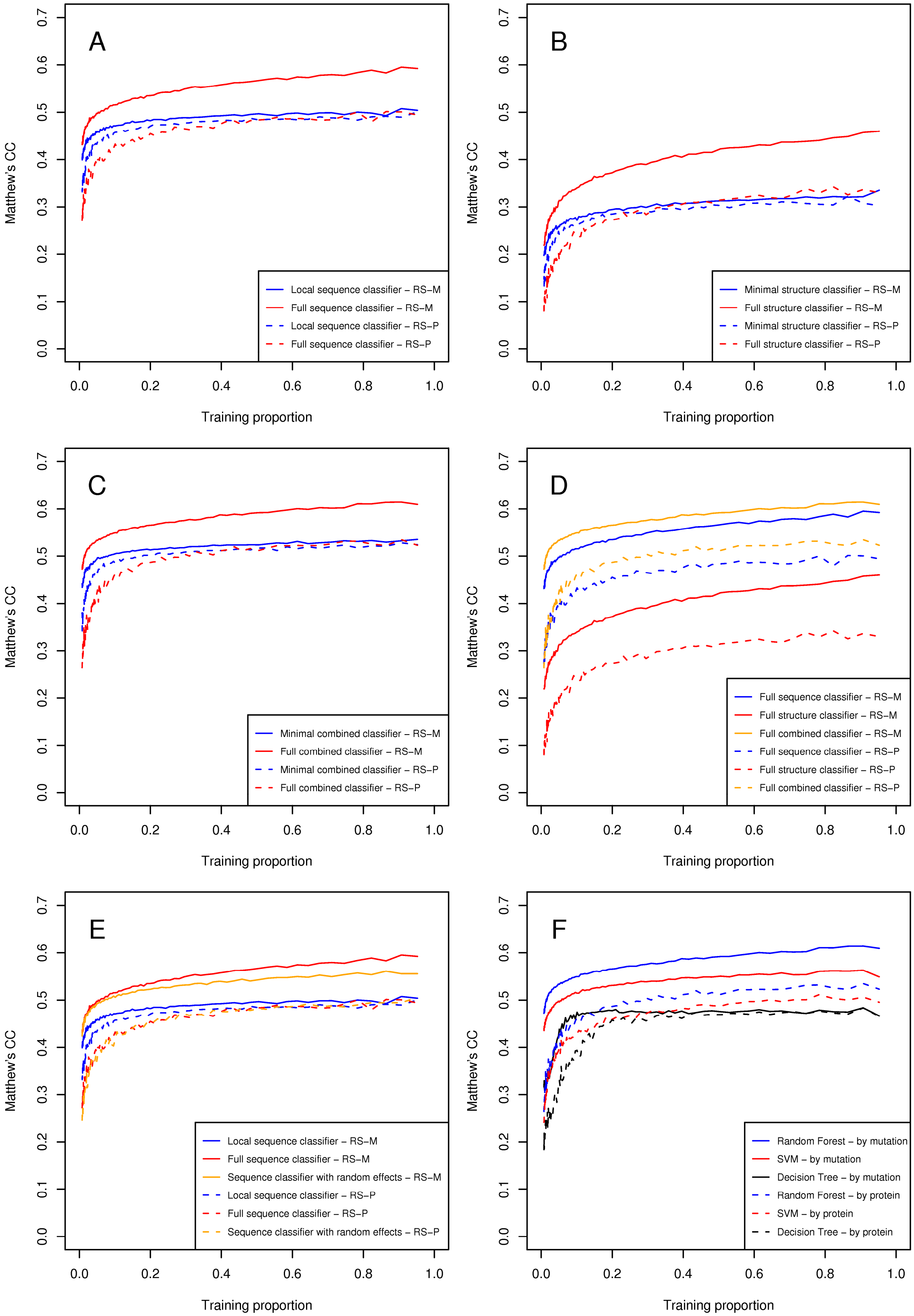


Figure 1: The Matthew's Correlation Coefficient (MCC) of the various classifiers is displayed

as a function of the proportion of the data used for training. For validating a machine learning model, the data is partitioned into a set for training the method, which is validated using the remainder. Splits are performed either at the mutation level (RS-M) or at the protein level (RS-P); they are repeated 30 times and the average MCC is reported. Several comparisons are performed: (A) full sequence information versus local sequence information, (B) the full structure model incorporating 10 pseudo-energies versus the minimal structure model using only the best pseudo-energy term, (C) the full combined classifier and the minimal classifier, (D) a comparison of the full sequence, structure and combined models, (E) the whole protein conservation scores versus protein level random factors and (F) a comparison of three machine learning methods using the full combined descriptors.

3. Summary and Outlook

The analysis of alternative structures shows that roughly half of the positional variation of the atoms in equivalent amino acids observed in independently solved structures can be predicted. Descriptors capturing the quality of the experimental data, the fit of the structure to the experimental data and geometric features of the amino acid residues can be combined in a log linear model for predicting the expected variation at a given site. Between pairs of protein structures, there is considerable variability in the degree to which the structural variation can be characterised, showing that despite rigorous efforts for filtering the alternative structures used here, numerous as yet to be determined confounders still upset the statistical model. Using the predicted variation and the model's standard error, the absolute values for the variation can be converted to Z-scores, reflecting the significance, rather than the absolute magnitude of variation. This method is applied to analyse the impact of point mutations on structures. Here, single point mutant pairs are classed by their evolutionary favourability; based on the difference in likelihood of observing each of the residues at the mutation site in an alignment of homologous proteins, mutant structure pairs were classed as conservative and non-conservative. The variation of the sites in the vicinity of the mutation is insignificant, when using the raw measure of deviation. Applying the normalisation step, it emerges that the non-conservative mutations vary more highly than the conservative ones. This effect is exacerbated, when one analyses only proteins for which the overall variability can be well predicted. For these proteins, the raw values of variability would lead one to believe that the conservative mutation vary more than the non-conservative ones, whereas the Z-score reverses this effect.

The practical applications of this model are twofold. (1) Z-scores are potentially more sensitive to detecting an excess of variability between structures; besides being applied to single point mutants, they could be used to infer structural movement in other cases, in which small but significant changes can occur, such as ligand binding. (2) The use of Z-scores rather than raw scores for benchmarking protein structure prediction methods may improve the accuracy estimates, particularly in the special case of point mutation prediction, which aims to model structural changes induced by single amino acid alterations. By knowing the expected level of noise in a structure, weighting schemes can be employed to take this into account during assessment of prediction accuracy.

The analysis of methods for protein structure prediction during the CASP-7 experiment shed light on the state of the art of current prediction tools. For the purpose of improving structural coverage for protein sequences, 3 main findings are of particular importance. (1) Protein structure prediction methods are still heavily dependent on structural templates. The quality of the best models decreases steadily with increasing target difficulty and even methods which use rigorous structural sampling do not achieve near-native models. The fact that physics based methods are unlikely to make accurate predictions of protein structures in the absence of templates confirms the relevance of homology modelling as a tool for protein structure prediction. (2) While current modelling tools may not be able to solve the *ab initio* protein folding problem, their accuracy is vastly improved over basic protocols, which rely on template identification using PSI-BLAST and copying the template coordinates. It must be added though, that the PSI-BLAST method used here had a significant handicap, in that it was decided that missing loops would not be built. While this allows for a fair comparison by ensuring that only information from the template found by PSI-BLAST is used, it is an artificial restriction, which may underestimate the accuracy of PSI-BLAST based tools. (3) The best performing automated methods were not necessarily the most computationally expensive ones. The program HHSearch is a comparatively fast method, which uses a HMM-HMM matching method to find the best templates. It was only exceeded by the Zhang server, which uses more advanced fold-recognition and structural sampling methods to predict protein structures. In all, it is apparent that the protein structure prediction community still has a lot to accomplish if the protein folding problem is to be solved, but for the purpose of increasing structural coverage by homology modelling, a number of new, powerful tools are now available.

A few potential improvements to the assessment protocols used are worth discussing. In previous CASP experiments, emphasis has been placed on accuracy measures such as GDT-TS and AL0, which are reliant on structure superposition of the alpha carbon backbones. These methods neglect the importance of correct contacts, which arguably is very important if the predicted structure is to be of any practical use in interpreting biological phenomena or for generating hypotheses for experimental testing. The score used here, HB-score, contains information about correctly predicted side-chain interactions and thus provides a goodness measure for evaluating the number of correctly predicted contacts. Its drawback is that the secondary structure hydrogen bonds may numerically outweigh those involving side chains. Not only do they tend to be more numerous, but it is comparatively easy to predict helical secondary structures compared to the topology of the tertiary structure. Furthermore, the number of hydrogen bonds varies widely between structures making the score susceptible to variation. Two modifications could be used to improve the assessment of contact accuracy. First, a sequence separation criterion, such as a minimal distance of 6 residues along the linear amino acid sequence, would eliminate local contacts stemming from helical secondary structures. Second, a score based on the contacts between all atoms, rather than just between hydrogen bonding partners, would have the advantage of being independent of the sequence composition. These two modifications would eliminate local contact bias and weight all non-local contacts equally.

Finally, a classification scheme is presented here for predicting the effects of single point mutations. The Swiss-Prot human variants dataset used here is manually annotated using information from the scientific literature. Nevertheless, it cannot be excluded that some are misclassified, which in turn could limit the maximal accuracy of classifiers trained on it. For instance SNPs may be merely linked to a phenotype, rather than being causative of it; close proximity on the chromosome of a SNP to the true causative mutation might lead to the SNP to be annotated as disease causing, when its molecular phenotype is actually neutral. Conversely, mutations which do have a molecular phenotype may be classed as neutral polymorphisms, if they perform a function which is either not essential or can be compensated for. Bromberg and Rost[124] have argued that their method is limited by the quality of the experimental data used for training; it is conceivable that improved data will yield better classification models. The upper limit of mutation phenotype prediction accuracy is in the region of a 75-80%; this is observed for many current methods and is presumably due in part to the limits of data accuracy.

For the prediction of mutant phenotypes, sequence based information is very powerful, the most informative descriptors being derived from position specific scoring matrices. The value of whole protein conservation scores is contentious. While they allow for great increases in accuracy, when using mutation level cross validation, the more rigorous protein level validation shows that this information is not transferable between proteins. Structural data by itself can be highly informative, but does not appear to be very robust, as it is strongly dependent on the proportion of data points used in training. However, it can complement sequence data to a noticeable degree. The usual interpretation of sequence information relies on the assumption that the vast majority of changes in evolution are neutral, with the remainder being positively selected for. However, it is known that deleterious mutations can be compensated for by changes at neighbouring sites, and so variation in sequence alignments can falsely imply mutational tolerance. Structural information may complement this shortcoming, by determining the structural plasticity of a site.

Prospects for future work are the inclusion of quaternary structure information as well as biochemical pathway information. The misclassification rate for proteins which are natively multimeric or bind other biomolecules, will be higher using the current method than need be due to the lack of information on the interactions with binding partners. The development of homology modelling methods for predicting quaternary structure would be of obvious use in this respect, and would conceivably lead to even better results in classification.

4. References

1. Anfinsen, C.B. Principles that govern the folding of protein chains. *Science (New York, N.Y.)* **181**, 223-30 (1973).

2. Chothia, C. & Lesk, A.M. The relation between the divergence of sequence and structure in proteins. *The EMBO journal* **5**, 823-6 (1986).

3. Murzin, A.G. Biochemistry. Metamorphic proteins. *Science (New York, N.Y.)* **320**, 1725-6 (2008).

4. Tuinstra, R.L. et al. Interconversion between two unrelated protein folds in the lymphotactin native state. *Proceedings of the National Academy of Sciences of the United States of America* **105**, 5057-62 (2008).

5. Meier, S. et al. Continuous molecular evolution of protein-domain structures by single amino acid changes. *Current Biology: CB* **17**, 173-8 (2007).

6. Dyson, H.J. & Wright, P.E. Intrinsically unstructured proteins and their functions. *Nature Reviews. Molecular Cell Biology* **6**, 197-208 (2005).

7. Fink, A.L. Natively unfolded proteins. *Current Opinion in Structural Biology* **15**, 35-41 (2005).

8. Ward, J.J. et al. Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *Journal of Molecular Biology* **337**, 635-45 (2004).

9. Dyson, H.J. & Wright, P.E. Coupling of folding and binding for unstructured proteins. *Current Opinion in Structural Biology* **12**, 54-60 (2002).

10. Levinthal, C. Are there pathways for protein folding? *Journal de Chimie Physique et de Physico-Chimie Biologique* **65**, 44-45 (1968).

11. Dill, K.A. & Chan, H.S. From Levinthal to pathways to funnels. *Nature Structural Biology* **4**, 10-19 (1997).

12. Pace, C.N. Conformational stability of globular proteins. *Trends in Biochemical Sciences* **15**, 14-7 (1990).

13. Lehmann, M. & Wyss, M. Engineering proteins for thermostability: the use of sequence alignments versus rational design and directed evolution. *Current Opinion in Biotechnology* **12**, 371-5 (2001).

14. Jaenicke, R. & Böhm, G. The stability of proteins in extreme environments. *Current Opinion in Structural Biology* **8**, 738-48 (1998).

15. DePristo, M.A., Weinreich, D.M. & Hartl, D.L. Missense meanderings in sequence space: a biophysical view of protein evolution. *Nature Reviews. Genetics* **6**, 678-87 (2005).

16. Hubbard, S.J. & Argos, P. Evidence on close packing and cavities in proteins. *Current Opinion in Biotechnology* **6**, 375-81 (1995).

17. Chen, J. & Stites, W.E. Packing is a key selection factor in the evolution of protein hydrophobic cores. *Biochemistry* **40**, 15280-9 (2001).

18. Dahiyat, B.I. & Mayo, S.L. Probing the role of packing specificity in protein design. *Proceedings of the National Academy of Sciences of the United States of America* **94**, 10172-7 (1997).

19. Wolfenden, R. Experimental measures of amino acid hydrophobicity and the topology of transmembrane and globular proteins. *The Journal of General Physiology* **129**, 357-62 (2007).

20. Dill, K.A. et al. The protein folding problem. *Annual Review of Biophysics* **37**, 289-316 (2008).

21. Baker, E.N. & Hubbard, R.E. Hydrogen bonding in globular proteins. *Progress in Biophysics and Molecular Biology* **44**, 97-179 (1984).

22. Myers, J.K. & Pace, C.N. Hydrogen bonding stabilizes globular proteins. *Biophysical Journal* **71**, 2033-9 (1996).

23. Byrne, M.P. et al. Energetic contribution of side chain hydrogen bonding to the stability of staphylococcal nuclease. *Biochemistry* **34**, 13949-60 (1995).

24. Roseman, M.A. Hydrophobicity of the peptide C=O...H-N hydrogen-bonded group. *Journal of Molecular Biology* **201**, 621-3 (1988).

25. Yang, A.S. & Honig, B. Free energy determinants of secondary structure formation: I. alpha-Helices. *Journal of Molecular Biology* **252**, 351-65 (1995).

26. Yang, A.S. & Honig, B. Free energy determinants of secondary structure formation: II. Antiparallel beta-sheets. *Journal of Molecular Biology* **252**, 366-76 (1995).

27. Schueler, O. & Margalit, H. Conservation of Salt Bridges in Protein Families. *Journal of Molecular Biology* **248**, 125-135 (1995).

28. Gordon, D.B., Marshall, S.A. & Mayo, S.L. Energy functions for protein design. *Current Opinion in Structural Biology* **9**, 509-13 (1999).

29. Hendsch, Z.S. & Tidor, B. Do salt bridges stabilize proteins? A continuum electrostatic analysis. *Protein Science: A Publication of the Protein Society* **3**, 211-26 (1994).

30. Waldburger, C.D., Schildbach, J.F. & Sauer, R.T. Are buried salt bridges important for protein stability and conformational specificity? *Nature Structural Biology* **2**, 122-128 (1995).

31. Elcock, A.H. Prediction of functionally important residues based solely on the computed energetics of protein structure. *Journal of Molecular Biology* **312**, 885-96 (2001).

32. Xiao, L. & Honig, B. Electrostatic contributions to the stability of hyperthermophilic proteins. *Journal of Molecular Biology* **289**, 1435-1444 (1999).

33. Pace, C.N. et al. Conformational stability and activity of ribonuclease T1 with zero, one, and two intact disulfide bonds. *The Journal of Biological Chemistry* **263**, 11820-5 (1988).

34. Cooper, A. et al. Thermodynamic consequences of the removal of a disulphide bridge from hen lysozyme. *J Mol Biol* **225**, 939-43 (1992).

35. Williams, R.A., Mamotte, C.D. & Burnett, J.R. Phenylketonuria: an inborn error of phenylalanine metabolism. *The Clinical Biochemist. Reviews / Australian Association of Clinical Biochemists* **29**, 31-41 (2008).

36. Collins, F. Cystic fibrosis: molecular biology and therapeutic implications. *Science* **256**, 774-779 (1992).

37. Sidransky, D. & Hollstein, M. Clinical implications of the p53 gene. *Annual Review of Medicine* **47**, 285-301 (1996).

38. Guerois, R., Nielsen, J.E. & Serrano, L. Predicting Changes in the Stability of Proteins and Protein Complexes: A Study of More Than 1000 Mutations. *Journal of Molecular Biology* **320**, 369-387 (2002).

39. Pakula, A.A., Young, V.B. & Sauer, R.T. Bacteriophage lambda cro mutations: effects on activity and intracellular degradation. *Proceedings of the National Academy of Sciences of the United States of America* **83**, 8829-33 (1986).

40. Erlandsen, H. & Stevens, R.C. The structural basis of phenylketonuria. *Molecular Genetics and Metabolism* **68**, 103-25 (1999).

41. Erlandsen, H. et al. Structural studies on phenylalanine hydroxylase and implications toward understanding and treating phenylketonuria. *Pediatrics* **112**, 1557-65 (2003).

42. Ellis, R.J. Macromolecular crowding: an important but neglected aspect of the intracellular environment. *Current Opinion in Structural Biology* **11**, 114-9 (2001).

43. Goldberg, A.L. Protein degradation and protection against misfolded or damaged proteins. *Nature* **426**, 895-9 (2003).

44. Collins, T. et al. Activity, Stability and Flexibility in Glycosidases Adapted to Extreme Thermal Environments. *Journal of Molecular Biology* **328**, 419-428 (2003).

45. Müller, C.W. et al. Adenylate kinase motions during catalysis: an energetic counterweight balancing substrate binding. *Structure (London, England: 1993)* **4**, 147-56 (1996).

46. Daniel, R.M. et al. The role of dynamics in enzyme activity. *Annual Review of Biophysics and Biomolecular Structure* **32**, 69-92 (2003).

47. Mattevi, A., Bolognesi, M. & Valentini, G. The allosteric regulation of pyruvate kinase. *FEBS Letters* **389**, 15-9 (1996).

48. Hochstrasser, M. Ubiquitin, proteasomes, and the regulation of intracellular protein degradation. *Current Opinion in Cell Biology* **7**, 215-23 (1995).

49. Hochstrasser, M. Ubiquitin-dependent protein degradation. *Annual Reviews in Genetics* **30**, 405-439 (1996).

50. Schwartz, A.L. & Ciechanover, A. Targeting Proteins for Destruction by the Ubiquitin System: Implications for Human Pathobiology. *Annual Review of Pharmacology and Toxicology* doi:10.1146/annurev.pharmtox.051208.165340 (2008).

51. Ellis, R.J. Macromolecular crowding: obvious but underappreciated. *Trends in Biochemical Sciences* **26**, 597-604 (2001).

52. Chiti, F. & Dobson, C.M. Protein misfolding, functional amyloid, and human disease. *Annual Review of Biochemistry* **75**, 333-66 (2006).

53. Honig, B. & Nicholls, A. Classical electrostatics in biology and chemistry. *Science (New York, N.Y.)* **268**, 1144-9 (1995).

54. Havranek, J.J., Duarte, C.M. & Baker, D. A simple physical model for the prediction and design of protein-DNA interactions. *Journal of Molecular Biology* **344**, 59-70 (2004).

55. Ashworth, J. et al. Computational redesign of endonuclease DNA binding and cleavage specificity. *Nature* **441**, 656-9 (2006).

56. Gutteridge, A. & Thornton, J.M. Understanding nature's catalytic toolkit. *Trends in Biochemical Sciences* **30**, 622-9 (2005).

57. Cheng, S.H. et al. Defective intracellular transport and processing of CFTR is the molecular basis of most cystic fibrosis. *Cell* **63**, 827-34 (1990).

58. Kruglyak, L. & Nickerson, D.A. Variation is the spice of life. *Nature Genetics* **27**, 234-6 (2001).

59. Brookes, A.J. The essence of SNPs. *Gene* **234**, 177-86 (1999).

60. Eyre-Walker, A. & Keightley, P.D. High genomic deleterious mutation rates in hominids. *Nature* **397**, 344-7 (1999).

61. Clegg, R.M. Fluorescence resonance energy transfer. *Current Opinion in Biotechnology* **6**, 103-110 (1995).

62. Wu, P. & Brand, L. Resonance energy transfer: methods and applications. *Analytical Biochemistry* **218**, 1-13 (1994).

63. Brunger, Axel T The free R value: a novel statistical quantity for assessing the accuracy of crystal structures. *Nature* **355**, 472-4 (1992).

64. Ringe, D. & Petsko, G.A. Study of protein dynamics by X-ray diffraction. *Methods in enzymology* **131**, 389-433 (1986).

65. Berman, H.M. et al. The Protein Data Bank. *Acta crystallographica. Section D, Biological crystallography* **58**, 899-907 (2002).

66. Eyal, E. et al. Protein side-chain rearrangement in regions of point mutations. *Proteins* **50**, 272-82 (2003).

67. Eyal, E. et al. MutaProt: a web interface for structural analysis of point mutations. *Bioinformatics (Oxford, England)* **17**, 381-2 (2001).

68. Feyfant, E., Sali, A. & Fiser, A. Modeling mutations in protein structures. *Protein Science: A Publication of the Protein Society* **16**, 2030-41 (2007).

69. Bordner, A.J. & Abagyan, R.A. Large-scale prediction of protein geometry and stability changes for arbitrary single point mutations. *Proteins* **57**, 400-13 (2004).

70. Bott, R. & Frane, J. Incorporation of crystallographic temperature factors in the statistical analysis of protein tertiary structures. *Protein Engineering* **3**, 649-57 (1990).

71. Graycar, T. et al. Engineered Bacillus lentus subtilisins having altered flexibility. *Journal of Molecular Biology* **292**, 97-109 (1999).

72. Kimura, M. Evolutionary Rate at the Molecular Level. *Nature* **217**, 624-626 (1968).

73. Smith, N.G.C. & Eyre-Walker, A. Adaptive protein evolution in Drosophila. *Nature* **415**, 1022-4 (2002).

74. Dayhoff, M.O., Schwartz, R.M. & Orcutt, B.C. In: Dayhoff, MO (Ed.), Atlas of Protein Sequence and Structure, vol. 5, Suppl. 3. *National Biomedical Research Foundation, Washington, DC* 345–352 (1978).

75. Henikoff, S. & Henikoff, J.G. Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences of the United States of America* **89**, 10915-9 (1992).

76. Yue, P. & Moult, J. Identification and analysis of deleterious human SNPs. *Journal of molecular biology* **356**, 1263-74 (2006).

77. Ng, P.C. & Henikoff, S. Predicting deleterious amino acid substitutions. *Genome research* **11**, 863-74 (2001).

78. Yip, Y.L. et al. Structural assessment of single amino acid mutations: application to TP53 function. *Human mutation* **27**, 926-37 (2006).

79. Valdar, W.S.J. Scoring residue conservation. *Proteins* **48**, 227-41 (2002).

80. Ng, P.C. & Henikoff, S. SIFT: Predicting amino acid changes that affect protein function. *Nucleic acids research* **31**, 3812-4 (2003).

81. Ng, P.C. & Henikoff, S. Predicting the effects of amino acid substitutions on protein function. *Annual review of genomics and human genetics* **7**, 61-80 (2006).

82. Ng, P.C. & Henikoff, S. Accounting for human polymorphisms predicted to affect protein function. *Genome Research* **12**, 436-46 (2002).

83. Sunyaev, S. et al. Prediction of deleterious human alleles. *Human molecular genetics* **10**, 591-7 (2001).

84. Sunyaev, S.R. et al. PSIC: profile extraction from sequence alignments with position-specific counts of independent observations. *Protein Engineering* **12**, 387-94 (1999).

85. Altschul, S.F. et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* **25**, 3389-402 (1997).

86. Göbel, U. et al. Correlated mutations and residue contacts in proteins. *Proteins* **18**, 309-17 (1994).

87. Hulo, N. et al. The 20 years of PROSITE. *Nucleic Acids Research* **36**, D245-9 (2008).

88. Finn, R.D. et al. The Pfam protein families database. *Nucleic Acids Research* **36**, D281-8 (2008).

89. Bairoch, A. & Apweiler, R. The SWISS-PROT protein sequence data bank and its new supplement TREMBL. *Nucleic Acids Research* **24**, 21-5 (1996).

90. Schwede, T. et al. SWISS-MODEL: An automated protein homology-modeling server. *Nucleic acids research* **31**, 3381-5 (2003).

91. Dobson, R.J. et al. Predicting deleterious nsSNPs: an analysis of sequence and structural attributes. *BMC Bioinformatics* **7**, 217 (2006).

92. Yue, P., Li, Z. & Moult, J. Loss of protein structure stability as a major causative factor in monogenic disease. *Journal of molecular biology* **353**, 459-73 (2005).

93. Altschul, S.F. et al. Basic local alignment search tool. *Journal of Molecular Biology* **215**, 403-10 (1990).

94. Moult, J. et al. Critical assessment of methods of protein structure prediction-Round VII. *Proteins* **69 Suppl 8**, 3-9 (2007).

95. Sunyaev, S., Ramensky, V. & Bork, P. Towards a structural basis of human non-synonymous single nucleotide polymorphisms. *Trends in genetics : TIG* **16**, 198-200 (2000).

96. Wang, Z. & Moult, J. SNPs, protein structure, and disease. *Human mutation* **17**, 263-70 (2001).

97. Wang, Z. & Moult, J. Three-dimensional structural location and molecular functional effects of missense SNPs in the T cell receptor Vbeta domain. *Proteins* **53**, 748-57 (2003).

98. Steward, R.E. et al. Molecular basis of inherited diseases: a structural perspective. *Trends in Genetics: TIG* **19**, 505-13 (2003).

99. Chasman, D. & Adams, R.M. Predicting the functional consequences of non-synonymous single nucleotide polymorphisms: structure-based assessment of amino acid variation. *Journal of molecular biology* **307**, 683-706 (2001).

100. Terp, B.N. et al. Assessing the relative importance of the biophysical properties of amino acid substitutions associated with human genetic disease. *Human mutation* **20**, 98-109 (2002).

101. Ferrer-Costa, C., Orozco, M. & de la Cruz, X. Characterization of disease-associated single amino acid polymorphisms in terms of sequence and structure properties. *Journal of molecular biology* **315**, 771-86 (2002).

102. Ferrer-Costa, C., Orozco, M. & de la Cruz, X. Sequence-based prediction of pathological mutations. *Proteins* **57**, 811-9 (2004).

103. Ferrer-Costa, C., Orozco, M. & de la Cruz, X. Use of bioinformatics tools for the annotation of disease-associated mutations in animal models. *Proteins* **61**, 878-87 (2005).

104. Brooks, B. et al. CHARMM: a program for macromolecular energy, minimization, and dynamics calculations. *Journal of computational chemistry* **4**, 187-217 (1983).

105. van Gunsteren, W.F. & Berendsen, H.J.C. Groningen Molecular Simulation (GROMOS) Library Manual. *Biomos, Groningen* (1987).

106. Cornell, W.D. et al. A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *Journal of the American Chemical Society* **117**, 5179-5197 (1995).

107. Zoete, V. & Meuwly, M. Importance of individual side chains for the stability of a protein fold: computational alanine scanning of the insulin monomer. *Journal of Computational Chemistry* **27**, 1843-57 (2006).

108. Pitera, J.W. & Kollman, P.A. Exhaustive mutagenesis in silico: multicoordinate free energy calculations on proteins and peptides. *Proteins* **41**, 385-97 (2000).

109. Sippl, M.J. Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins. *Journal of Molecular Biology* **213**, 859-83 (1990).

110. Zhou, H. & Zhou, Y. Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. *Protein Science: A Publication of the Protein Society* **11**, 2714-26 (2002).

111. Gilis, D. & Rooman, M. Stability changes upon mutation of solvent-accessible residues in proteins evaluated by database-derived potentials. *Journal of Molecular Biology* **257**, 1112-26 (1996).

112. Gilis, D. & Rooman, M. Predicting protein stability changes upon mutation using database-derived potentials: solvent accessibility determines the importance of local versus non-local interactions along the sequence. *Journal of Molecular Biology* **272**, 276-90 (1997).

113. Krishnan, V.G. & Westhead, D.R. A comparative study of machine-learning methods to predict the effects of single nucleotide polymorphisms on protein function. *Bioinformatics (Oxford, England)* **19**, 2199-209 (2003).

114. Breiman, L. et al. CART: Classification and Regression Trees. *Wadsworth: Belmont, CA* (1983).

115. Hu, J. & Yan, C. Identification of deleterious non-synonymous single nucleotide polymorphisms using sequence-derived information. *BMC Bioinformatics* **9**, 297 (2008).

116. Breiman, L. Random Forests. *Machine Learning* **45**, 5-32 (2001).

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

118. Barenboim, M. et al. Statistical geometry based prediction of nonsynonymous SNP functional effects using random forest and neuro-fuzzy classifiers. *Proteins* (2008).

119. Vapnik, V. Statistical Learning theory. (1998).

120. Karchin, R., Kelly, L. & Sali, A. Improving functional annotation of non-synonomous SNPs with information theory. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing* 397-408doi:15759645 (2005).

121. Karchin, R. et al. LS-SNP: large-scale annotation of coding non-synonymous SNPs based on multiple information sources. *Bioinformatics (Oxford, England)* **21**, 2814-20 (2005).

122. Moult, J. A decade of CASP: progress, bottlenecks and prognosis in protein structure prediction. *Current Opinion in Structural Biology* **15**, 285-9 (2005).

123. Yip, Y.L. et al. Retrieving mutation-specific information for human proteins in UniProt/Swiss-Prot Knowledgebase. *Journal of Bioinformatics and Computational Biology* **5**, 1215-31 (2007).

124. Bromberg, Y. & Rost, B. SNAP: predict effect of non-synonymous polymorphisms on function. *Nucleic acids research* **35**, 3823-35 (2007).

125. Kopp, J. et al. Assessment of CASP7 predictions for template-based modeling targets. *Proteins* **69 Suppl 8**, 38-56 (2007).

126. Zemla, A. LGA: A method for finding 3D similarities in protein structures. *Nucleic acids research* **31**, 3370-4 (2003).

127. McDonald, I.K. & Thornton, J.M. Satisfying hydrogen bonding potential in proteins. *Journal of molecular biology* **238**, 777-93 (1994).

128. Hubbard, S.J. & Thornton, J.M. NACCESS.

5. Acknowledgements

I would like to thank Prof. Torsten Schwede for giving me the opportunity to perform my PhD studies in his group at the Biozentrum and for his support during my time at the Biozentrum. I am also grateful to Prof. Olivier Michielin for agreeing to act as my thesis examiner.

I am indebted to Lorenza Bordoli for critically reading my thesis and providing constructive comments. I would like to thank my colleagues in the structural bioinformatics group for fruitful discussions and trading code and ideas, as well being generally good-humoured.

And last but certainly not least I would like to thank my parents and sister for their help and support.

Financially this work was supported by the SNF (grant number 3100-068327), the Swiss Institute of Bioinformatics and the Biozentrum Basel.