Graph-based Approaches to Protein Structure- and Function Prediction

Dissertation zur Erlangung des Grades
eines Doktors der Naturwissenschaften (Dr. rer. nat.)
am Fachbereich Mathematik und Informatik
der Freien Universität Berlin

von

Henning Stehr

Berlin
Mai 2011

Datum der Disputation:
30. September 2011

Gutachter:
Prof. Dr. Martin Vingron
PD Dr. Robert Preissner

獻給我的愛妻

Abstract

The three-dimensional structure of a protein is determined by the network of covalent and non-covalent interactions. An exact description of the governing forces requires to take into account quantum effects which makes the system too complex for most computational analyses. Depending on the application, a simpler representation can make such analyses feasible while still capturing the relevant aspects of the system. Here we explore a number of applications which are based on representing a protein as a graph of interacting residues. This representation has some conceptual advantages over an all-atom representation. It can be shown that the graph representation captures the three dimensional information up to an average accuracy of 2 Ångstrom C-alpha RMSD. The deviation stems from the fact that the network is equivalent to an ensemble of structures which satisfy the contact constraints. This can be used to represent some degree of flexibility. Furthermore, the representation makes it possible to apply algorithms from graph theory to common protein analysis problems such as structure alignment and structure prediction.

Part 1: Multiple Structure Alignment and the Sample Mean of Graphs. In contrast to the alignment of sequences, structure-based alignments allow us to look further back in evolutionary time when comparing proteins. For the pairwise case, it has been shown that graph-based methods yield very sensitive alignments which are able to detect remote evolutionary relationships. Here we extend this approach to the case of aligning multiple proteins represented as graphs. This gives rise to a mathematically rigorous definition of the optimal multiple alignment. We analytically derive that calculating an optimal alignment is equivalent to calculating the sample mean for a set of graphs. This sample mean theory for graphs has only very recently been developed and makes a number of powerful algorithms applicable to protein structure analysis. We propose a new multiple structure alignment algorithm based on the sample mean theory and compare its performance to current alternative methods. We show that our algorithm is more efficient than other graph-based algorithms while retaining the same advantages. We further show that the quality of the alignments are as good as other current methods when benchmarked on a large set of structural alignments.

Part 2: Consensus Prediction of Residue-Residue Contacts. Accurate prediction of the non-covalent interactions in a protein is equivalent to predicting the three-dimensional structure. Results from the CASP experiments show that current contact prediction methods are inferior to methods which attempt to predict the three dimensional structure directly. However, contact prediction can sometimes give complementary information and hence has been included in some of the top-performing structure prediction methods. Here we present a contact prediction method which is based on calculating the graph-mean of a number of input predictions to create a consensus prediction. We tested the method with CASP server predictions as inputs which are converted to contact graphs before calculating the sample mean. The parameters of the method are trained on targets from the CASP 7 experiment and evaluated on targets from CASP 8. Predicted contacts are evaluated in terms of accuracy and coverage compared to the native contacts. For 50% of the targets, our prediction is better than any of the input models. For 85% of the targets, our prediction is in the top 5% and in

all cases it is better than the median score. When compared to the individual methods from which input models were taken, our method predicts contacts more accuractely than any other method. This shows that even though many state-of-the-art methods already make use of consensus information for template picking and model selection, consensus information at the contact level can be further exploited to improve current prediction methods.

Part 3: The Structural Impact of Cancer-Associated Mutations in Oncogenes and Tumor Suppressors. Since the availability of high-throughput sequencing methods, the first complete cancer genomes have been published. By comparing the sequence variation between tumor cells and healthy tissue from the same patient, somatic mutations can be identified which are potentially linked with tumorigenesis. However, the consequence of the mutations on a structural and functional level remains to be characterized. We applied computational methods to study the effects of somatic mutations on known and predicted protein structures of well-known cancer genes. For each of $\approx$2000 mutations, we investigate surface accessibility, proximity to known functional sites, clustering of mutations within the structure, and stability change upon mutation. We obtain significant differences between mutations in oncogenes and tumor suppressors. While mutations in oncogenes tend to occur at the protein surface, are highly clustered and directly affect sites important for protein function, mutations in tumor suppressors tend to affect primarily protein stability. We also find that the alteration of oncogenic activity is often associated with mutations at ATP or GTP binding sites. With these results we can confirm and statistically validate the hypotheses for the gain-of-function and loss-of-function mechanisms of oncogenes and tumor suppressors, respectively. We further show that the differences in the mutational patterns can be used to predict for previously uncharacterized genes, identified in cancer sequencing studies, whether they will likely function as an oncogene or a tumor suppressor. This method can be a valuable tool in the analysis of the increasing amount of data that is being generated by the current cancer sequencing projects.

Summary: Starting from a theoretical result, the equivalence of structure alignment and calculating the sample mean of graphs, we present novel methods for the multiple alignment of protein structures and for the prediction of residue-residue contacts. These methods have direct applications in protein structure analysis. In the third part we apply structure prediction to the analysis of human disease mutations. The results confirm the gain of function vs. loss of function hypothesis of oncogenes and tumor suppressors, respectively and give rise to a method for predicting functional properties of cancer-associated proteins from their mutational patterns obtained from sequencing of cancer genomes.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

> *"That is to say, the polypeptide chain, once synthesized, should be capable of folding itself up without being provided with additional information; this capacity has, in fact, recently been demonstrated by Anfinsen in vitro for one protein, namely ribonuclease. If the postulate is true it follows that one should be able to predict the three-dimensional structure of a protein from a knowledge of its amino acid sequence alone. Indeed, in the very long run, it should only be necessary to determine the amino acid sequence of a protein, and its three-dimensional structure could then be predicted; in my view this day will not come soon, but when it does come the X-ray crystallographers can go out of business[...]."*

This quote by John Kendrew from his Nobel lecture held in 1962 nicely illustrates for how long the scientific community has been working on trying to solve the structures of proteins with computational means. Today, almost half a century later, although steady progress has been made, we are still busy working on the methods to make Kendrew's prediction come true, and to finally put crystallographers out of business.

Kendrew, together with Max Perutz, received the nobel price in chemistry in 1962 for his work on determining the first atomic structure of a protein using X-ray crystallography. This first structure, a sperm whale myoglobin extracted from a chunk of whale meat, was determined in 1959 only after several attempts to crystallize similar proteins had failed (Kendrew et al., 1960). In the same year, Perutz had succeeded in solving the structure of another protein, a horse hemoglobin (Perutz et al., 1960). Only two years after their work was published in 1960, they were jointly awarded the nobel prize.

The observation that myoglobin and hemoglobin, despite their different amino acid compositions, showed essentially the same tertiary structure lead Kendrew to comment that *"myoglobin possesses a structure the significance of which extends beyond a particular species and even beyond a particular protein"* (Perutz, 1997).

What seems obvious from today's perspective, that proteins which perform similar functions in different organisms and which possess similar, but distinct sequences share a common fold, was one of the fundamental discoveries of Perutz' and Kendrew's work.

In fact, Helen Scouloudi had already noted in an article published in Nature in 1959 that her results from two dimensional scattering experiments suggested that myoglobins from differents species seem to have the same tertiary structure (Scouloudi, 1959). The atomic-resolution structures by Perutz and Kendrew then proved not only this hypothesis, but also several other theoretical models, such as the shape of the alpha helix (Pauling et al., 1951).

So these pioneering experiments, 50 years ago, laid the foundations for two major problems which are still fundamental in computational structural biology today: determining a protein's structure given its sequence, and discovering similarites between proteins given their structures.

In this thesis, we will address these two problems with novel graph-based approaches. We will also show for an application, the analysis of cancer mutations, how knowledge of the structure can be used to gain insights into a protein's function in the cell and to understand how its malfunction can lead to human disease. Understanding these mechanisms is not only of fundamental scientific interest, but also has direct relevance for the development of new therapies.

In Chapter 1, we will review some of the methods for computational structure prediction which have been developed over the past decades and introduce the concept of modeling protein structures as graphs.

In Chapter 2, we will introduce a framework for protein structure comparison using graph theory and show how it can be algorithmically applied to the multiple structure alignment problem.

In Chapter 3, we will show how this framework can be used to predict a key aspect of a protein's structure, the intra-molecular interactions of amino acids.

Finally, in Chapter 4, we will use structure prediction to address a topic with direct consequences for understanding disease mechanisms: the structural impact of cancer-associated mutations.

1.1 Protein structure prediction

1.1.1 Physics-based methods

The study of protein folding goes back to the beginning of the 20th century. In 1910, Chick and Martin observed for the first time protein denaturation as a distinct process (Chick and Martin, 1910). Following further experiments, Anson and Mirsky noted that the process of denaturation was reversible (which by some was initially taunted as 'unboiling the egg') and that it involved free energy changes which were much smaller than those typically observed in chemical reactions (Anson and Mirsky, 1925). In 1929, Hsien Wu first hypothesized that the observed changes in soluability, enzymatic activity and chemical reactivity upon denaturation could be due to a simple conformational change of the protein chain, rather than a chemical modification (Wu, 1995; Edsall, 1995). This hypothesis was heavily disputed at first but later advocated by Mirsky and Pauling (Mirsky and Pauling, 1936) and finally, generally accepted.

The final confirmation of the principle concept of protein folding was achieved by Anfinson's famous experiments on ribonucleases (Anfinsen, 1973) in the 1960s for which he received the nobel price in 1972. His work established two very important results. First, that the information for the native structure is completely contained in the protein sequence. And second, that the native structure is the unique, stable and kinetically accessible minimum of the free energy. The latter, known as the *thermodynamic hypothesis* also implies that the same chain will always fold to the same native structure. These results can be seen as the foundations of computational structure prediction.

Of course, the principle possibility did not devise a method for finding the native structure. In his famous thought experiment, Cyrus Levinthal was contemplating about the question how the natural folding process finds the native conformation. He stated that, if a protein were to explore all possible conformations of its chain in order to find the lowest energy state, the folding process would take a longer timespan than the age of the universe (Levinthal, 1969). As a consequence, there must be a guiding process which directs the unfolded chain towards the folded conformation. This ideas is sometimes called the *folding funnel* model. This metaphor refers to the shape of the energy-landscape of the folding process, which, in three dimensions, can be visualized as a funnel-like shape. Even though the energy landscape will be locally rugged with many small local minima, the folding trajectory will basically follow a direction of steepest descent into the funnel (Dill and Chan, 1997). Even though Levinthal's 'paradox' can relatively easily be resolved by funnel-like models of the natural folding process (Zwanzig et al., 1992; Sali et al., 1994), it obviously has direct implications for computational structure prediction. It shows that a simple enumeration of states to find the lowest energy conformation is infeasible, a fact that is also reflected in later results that the folding problem in various subforms is NP-complete (Istrail and Lam, 2009).

Yet, with the emergence of accessible computer technology, the question arose, how the folding process could be simulated. The two principle approaches were, directly simulating physical behaviour (*Molecular Dynamics*), and stochastic simulations such

as the Metropolis Monte-Carlo method (Metropolis et al., 1953). The pioneering work on physical simulations was done by Levitt and coworkers (Levitt and Warshel, 1975; Levitt, 1976). It also established the important concept of *molecular mechanics forcefields* whose basic idea is to treat chemical bonds like mechanical springs and approximate the interaction energies with Newtonian-like mechanics (Burkert, 1982). An example for a typical molecular mechanics energy function is the equation for the stability given in Figure 4.4 in Chapter 4. This concept proved to be very fruitful to make simulations of whole proteins feasible and lead to many important results on molecular properties and folding behaviour (Duan and Kollman, 1998; Snow et al., 2002). In the following decades, much research effort has been invested in improved versions of molecular dynamics and related simulation techniques (Levitt and Sharon, 1988; Sugita and Okamoto, 1999; Leach, 2001). But despite these efforts and an increase in computing power of many orders of magnitude, the determination of a correctly folded structure for typical sized protein using these techniques remains infeasible. This is partly due to the still insufficient simulation times that can be achieved and partly due to the insufficient accuracy of the energy functions.

At the time of the first molecular dynamics simulations in the early 1970s, the number of experimentally resolved structures using X-ray crystallography had grown to about 20 (Berman, 2008). To keep the information about these structures in a common format and to make data sharing across different labs easier, the *Protein Data Bank* (PDB) was established at the Brookhaven National Laboratory (Hamilton, 1971). Previously, the data had been shared between labs in the form of punch cards with one card for each atom, so that for the structure of Myoglobin about 1000 cards were needed (Berman, 2008). The PDB was a premier example of the usefulness of central repositories for biological data. It enabled two important methodological advancements, the development of knowledge-based energy functions (Tanaka and Scheraga, 1976; Miyazawa and Jernigan, 1985; Sippl, 1995) and the field of template-based structure prediction.

1.1.2 Homology modeling

An alternative approach to simulations based on first principles is to exploit the data from databases of known protein structures. These approaches are usually called *knowledge-based* methods, and the most common one is *homology modeling*. This method is based on the observation that structure is more conserved than sequence, or in other words, proteins with similar sequences tend to have a very similar structure. So, to model the structure of one protein, one can use the information about the structure of a suitable homolog. The first homology model of an α-lactalbumin (an actual physical model made from wire), was already built in 1969 based on the crystal structure of lysozyme (Browne et al., 1969). Chothia and Lesk then did a quantitative analysis of the correlation between sequence- and structure divergence which is regarded as the foundation of homology modeling (Chothia and Lesk, 1986).

The typical procedure for predicting the structure of a protein (called the *target*), based on a known structure (the *template*), by homology modeling is as follows:

Input: The target sequence

1. Template identification
2. Target to template alignment
3. Model building
4. Modeling of loops and side chains

Output: The predicted structure of the target protein

The simplest way to find a template is to do a sequence search in a database of known structures (usually the PDB) with methods such as BLAST (Altschul et al., 1990) or Smith-Waterman alignment (Smith and Waterman, 1981). This works well for templates with high sequence similarity to the target. Template identification was greatly improved by the introduction of profile based methods such as PSI-BLAST (Altschul et al., 1997). They include evolutionary information by first building a multiple alignments of sequences similar to the target, then deriving a profile of amino acid frequencies from the columns of the alignment and finally, searching the database using the profile. This way, templates with less obvious sequence identities can be found (Park et al., 1998). In the last few years, several improved profile search methods have been developed. For example, the HHPred method, which generates profiles based on Hidden Markov Models for both the query and target sequences. It is able to detect very weak homology relationships (Hildebrand et al., 2009a). Once a template has been indentified, the target needs to be aligned to the template. This can either be done by simple sequence alignment or by sequence-to-structure alignment, which will be described in the following section on threading. The next step is the actual model building. One of the most popular model building program today is MODELLER (Sali and Blundell, 1993). It derives spatial constraints (distances and angles) from the template structure, assuming that amino acids which are in close distance in the template should also be close in the target model. A minimization method then finds conformations which satisfy the constraints derived from the template, and the stereochemical constraints of the target sequence. Another approach, used by the SWISS-MODEL method (Schwede et al., 2003), is to first identify a conserved core and then build up the model from small rigid-body parts of the template structure. A third type of approach is called 'artifical evolution' and basically proceeds by mutating the template structure, one amino acids at a time, followed by an energy minimization after every step (Petrey et al., 2003). Even for closely related structures, it is often advisable to model loop regions seperately, because they tend to be more diverged than the rest of the structure, and because they are often functionally relevant (Fiser et al., 2000). Loop modeling has been done by either taking loop conformations from known structures (Jones and Thirup, 1986; Michalsky et al., 2003; Hildebrand et al., 2009b) or by using ab-initio methods (Moult and James, 1986; Fiser and Sali, 2003; Jacobson et al., 2004; Soto et al., 2008; Mandell et al., 2009). Side-chain modelling is usually done by using so-called *rotamer libraries*, which contain discrete, frequently occuring side-chain conformations (Ponder and Richards, 1987b; Dunbrack and Karplus, 1993).

The side-chain packing problem then becomes a combinatorial optimization problem which is approached by either exact methods, such as dead-end-elimination (Desmet et al., 1992), or stochastic sampling methods (Holm and Sander, 1991).

For targets where a template with high sequence similarity can be found, homology modeling remains the most reliable structure prediction method. It has been shown that above 50% sequence identity, homology models are generally very accurate. Between 30% and 50% models are often correct but contain errors in loop regions and in side chain conformations. Below 30% (the so called 'twilight zone'), homology modeling becomes more and more unreliable (Rost, 1999; Baker and Sali, 2001).

1.1.3 Fold recognition

The aim of fold recognition methods is to find a structural template even if no detectable sequence similarity exists. This is based on the observation that there are much less distinct folds than sequences and many unrelated sequences fold into a similar structure (Wang, 1996, 1998; Zhang and DeLisi, 1998). By searching through the database of known structures, fold recognition methods seek to find a structure which is 'compatible' with the target sequence and hence which the target protein is likely to adopt. The idea of threading was introduced by Bowie and Eisenberg in 1991 (Bowie et al., 1991). The term *threading* first appeared in an article by Jones and coworkers in 1992 (Jones et al., 1992). Their idea was to 'thread' a sequence through each of several known structures to find a favourable conformation as illustrated in Figure 1.1. Each possible conformation is evaluated by a scoring function. Bowie et al. used

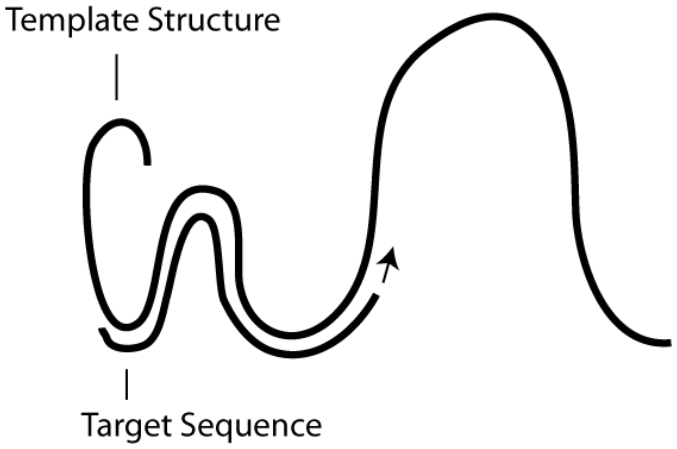

Figure 1.1: Illustration of protein threading. The target sequence is moved along a template structure to find a conformation the sequence 'likes' to fold into. This is evaluated by a scoring function which measures sequence-to-structure compatibility. Because in the general case, gaps are allowed between target and template, threading methods are also more generally called *sequence-to-structure alignment* methods.

the structural environment of each residue, such as secondary structure and solvent accessibility, to assess whether a particular amino acids 'likes' a given position in the

template structure. This has the advantage that dynamic programming can be used to find the best sequence-to-structure alignment. Jones et al. introduced pairwise interaction terms into the scoring function which improves the ability to detect the correct folds but also increases the computational complexity of the problem. It has been shown that the general threading problem with pairwise scoring and allowing for arbitrarily sized gaps in the alignment is NP-complete (Lathrop, 1994). While the original metaphor of 'threading' a sequence through a structure is nice to illustrate the original idea, the term is now being used for any method attempting to find a sequence-to-structure alignment for fold recognition. The methods differ in the search strategy for finding the best alignment and in the scoring functions being used. Apart from the dynamic programming methods already mentioned (Bowie et al., 1991; Torda et al., 2004), other search strategies include heuristic methods (Jones et al., 1992; Godzik et al., 1992; Westhead et al., 1995; Flöckner et al., 1995) and exact methods based on branch and bound (Lathrop and Smith, 1994; Xu and Xu, 2000), integer linear programming (Xu et al., 2003) and tree decomposition (Xu et al., 2005).

For the scoring functions being used, a general requirement is that they have to be reasonably fast to evaluate, because many conformations for each of the structures in the template library have to be screened. Also, the energy should not be too sensitive to the exact atomic positions. Typical physical energy functions do not fulfill these requirements. Therefore, threading has been a domain of residue-based empirical energy functions. The original simple pairwise potentials (Tanaka and Scheraga, 1976; Miyazawa and Jernigan, 1985; Sippl, 1995) were improved by including distance dependent terms (Zhou and Zhou, 2002), bond type specific terms (Nishikawa and Matsuo, 1993) and orientation-dependent terms (Buchete et al., 2004; Miyazawa and Jernigan, 2005).

Recently, the borders between the different types of template based methods have become blurred because sequence search methods are getting better at finding remote homologies which were previously the domain of threading. At the same time, ideas from threading methods have been incorporated into the more generally applicable fragment assembly methods which are described in the following section.

1.1.4 Fragment assembly

A limitation inherent in fold recognition approaches is that the fold adopted by the protein has to exist in the template database. The *fragment assembly* methods, developed since the late 1990s, set out to overcome this limitation. They make use of small fragments from known structures to build up models which do not necessarily need to have a previously known fold. In a first step, they collect fragments which parts of the target sequence are likely to fold into and then assemble the fragments into a final model. This idea has also been given the name *mini-threading*. The approach was encouraged by results that short six-residue fragments mostly fold into one of about 100 structural classes (Unger et al., 1989) and that protein backbones can, in principle, be built from fragments of other proteins (Kolodny et al., 2002; Jones and Thirup, 1986; Zhang and Skolnick, 2005). The FRAGFOLD method first proved the applicability of this approach to predict structures without a template in the PDB (Jones,

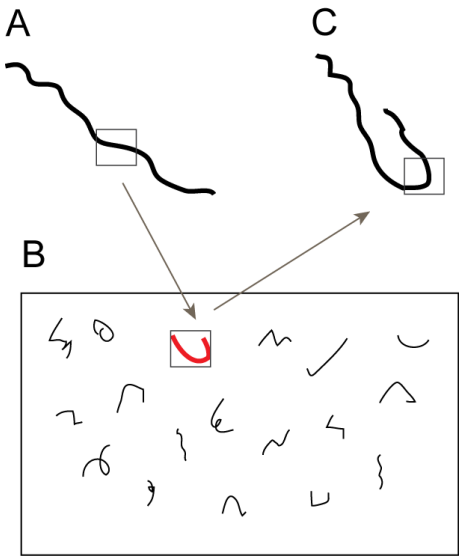

Figure 1.2: Illustration of a sampling step in the Rosetta fragment-assembly method (Simons et al., 1997). The target protein is initially represented as an unfolded chain (A). In every step, a position is chosen at random in the target chain. The fragment at that position is then replaced by a compatible fragment from the fragment-library (B). The new conformation (C) is evaluated with a scoring function. Depending on the score of the new conformation, the step is either accepted or rejected. This procedure is iterated until the chain is folded.

1997; Jones et al., 2005). The most successful method, according to the Casp rankings over the last years (Vincent et al., 2005; Jauch et al., 2007; Ben-David et al., 2009), has been the Rosetta method developed by the group of David Baker (Simons et al., 1997). It is a good example for illustrating how ideas from folding simulations and template based modeling could be successfully combined. The original Rosetta method uses a fragment library of nine-residue peptides with associated sequence propensities. It proceeds in five steps as illustrated in Figure 1.2. The target protein is initially represented as an unfolded chain (A).

1. Select a position p in the target chain at random.
2. Replace the nine-residue fragment f around p with a fragment from the library (B) which has a high propensity for the sequence of f. This creates a new conformation of the target chain (C).
3. Evaluate the new conformation using a course-grained empirical energy function.
4. Accept or reject the new conformation depending on the improvement in energy.
5. Repeat from step 1 until a stable conformation has been reached.

The scoring function used in step 3 contains both physical energy terms and empirical terms derived from known structures. It is tuned to distinguish large conformational changes, and therefore emphasizes global properties, like packing density and secondary structure formation, rather than atomic details, such as exact bond angles. If the energy of the new structure is improved, the move is accepted in step 4 with a

high probability. The bigger the improvement in energy, the higher the probability of acceptance. This probabilistic search ensures, that the simulation can escape local minima, but in general, will favor moves towards the minimum. Steps 1-5 are iterated according to a simulated annealing scheme (Kirkpatrick et al., 1983) such that, as the structure folds, less and less dramatic changes are allowed. Rosetta uses the strategy outlines above to generate many initial models, out of which a final model is picked, and further refined using all-atom refinement techniques.

1.1.5 Recent developments

This approach has worked well to predict the structures of many proteins, which had no structural templates in the PDB, with high accuracy (Qian et al., 2007; Raman et al., 2008; Jauch et al., 2007). With numerous improved methods currently being developed (Zhang and Skolnick, 2004; Fujitsuka et al., 2006; Wu et al., 2007; Ben-David et al., 2009), fragment assembly currently represents the most promising avenue for generally applicable protein structure prediction methods. The next breakthroughs can likely be expected from the integration of computational with experimental techniques such as NMR or Small Angle X-ray Scattering (Shen et al., 2008; Raman et al., 2010b,a; Schneidman-Duhovny et al., 2011). If such experimental methods can be tuned to deliver constraints for the computational sampling process in a high-throughput fashion, it seems possible that structure determination can start to close the gap between the number of known proteins and those with resolved structures.

Other areas of active research are the modeling of whole complexes (Alber et al., 2007; Förster et al., 2010) and, based on the insights gained from the work on structure prediction, the computational design of novel protein folds and enzymatic reactions which have not previosuly been observed in nature (Kuhlman et al., 2003; Röthlisberger et al., 2008; Jiang et al., 2008; Khersonsky et al., 2011).

1.2 Graph representation of protein structures

1.2.1 Ceci n'est pas une protéine - On the role of models

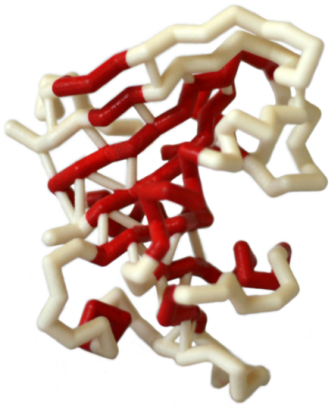

Figure 1.3: Dihydrofolate reductase

Figure 1.3 shows the picture of a protein. Or doesn't it?

Mechanisms at the level of molecules and atoms are governed by quantum laws. Since quantum laws are not only quite difficult to deal with mathematically, but also rather unintuitive or as Einstein put it "*a hopeless mess*", it is quite a relief that in our daily life, the approximations Newton proposed as the "laws of mechanics" are quite useful to get around. Now an even greater relief may be that even for proteins, a mechanistic description, where atoms can be imagined as spheres and bonds as springs, turns out to be quite useful in many circumstances.

Another problem we face when we study proteins, is that they are too small for us to see. The resolution of even the best light microscope is not sufficient to study individual proteins in detail. So what we need, is some way to indirectly measure certain properties of the system and then make the measurements visible for us to see. So what the early protein researchers did is to build models from paper, wire or clay.

What both of these examples, the sphere-and-spring model and the paper-protein model show, is that it is useful to have some mental image which we can relate to. It is no coincidence that protein research has been closely linked to the development of computer graphics. To visualize the three dimensional structure was an obvious use case for evolving computer graphics systems and molecular labs were early adopters of such systems (Levinthal, 1966; Ripka, 1986).

Figure 1.4 shows some typical depictions of proteins. They differ in that they highlight different molecular properties, for example, the distribution of electron density or the architecture of secondary structure elements. What the models have in common, is that they do not explain all of the properties of the system. We choose the models according to the properties we seek to capture and according to how well these are represented by the model. This characterization of a model is not restricted to graphical visualizations. In a more abstract sense, any description, if textual, mathematical or graphical, is a

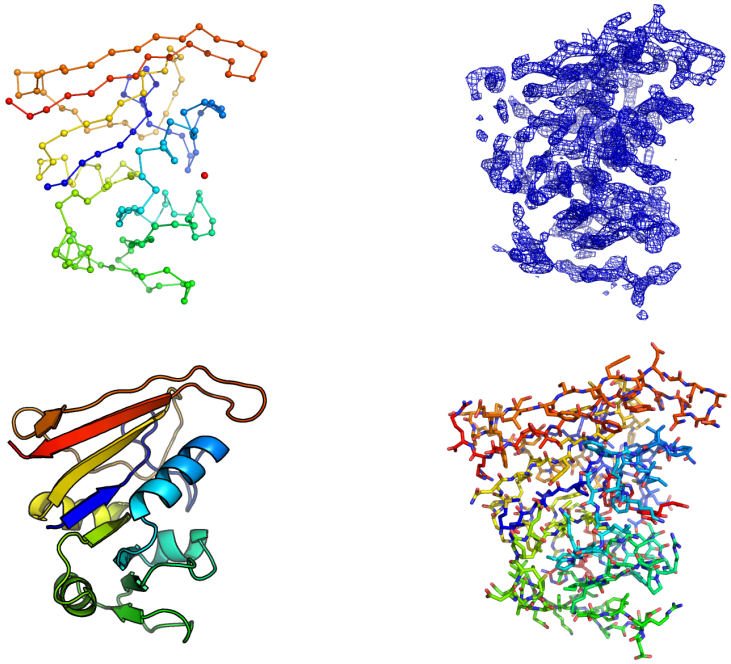

Figure 1.4: Protein depictions highlighting different molecular properties. The protein shown is Dihydrofolate reductase from E.coli (PDB code: 1dre). Top left: Backbone trace representation with C-alpha atoms shown as spheres. Top right: Contour plot of an electron density map as obtained by X-ray crystallography. Bottom left: Cartoon representation as known from text book examples. Bottom right: Heavy-atom representation with bonds shown as sticks.

model that describes certain aspects of the system under investigation. In that sense, it is closely related to the concept of a scientific theory (Freudenthal, 1961).

What we need to keep in mind is that a model can be more or less useful for a certain application but it will never capture the system in its entirety.

1.2.2 Residue interaction graphs and contact maps

The folded structure of a protein is held together by the network of covalent and non-covalent interactions of its atoms. The exact description of the governing forces requires to take into account quantum effects which makes the system too complex for most computational analyses. A simpler representation can make such analyses feasible while still capturing the aspects of the system which are relevant for the application. One particular way, which we explore in this thesis, is to represent the protein as a *residue-interaction-graph* or *contact graph*. That is, a graph where the nodes represent the protein's residues and two nodes are connected by an edge if the two residues are in contact in the structure.

Contact is usually defined as the distance between two representative atoms, for example the C-alpha atoms of the respective residues, falling below a *distance threshold*, for example 8Å. Other contact definitions which can be found in the literature include

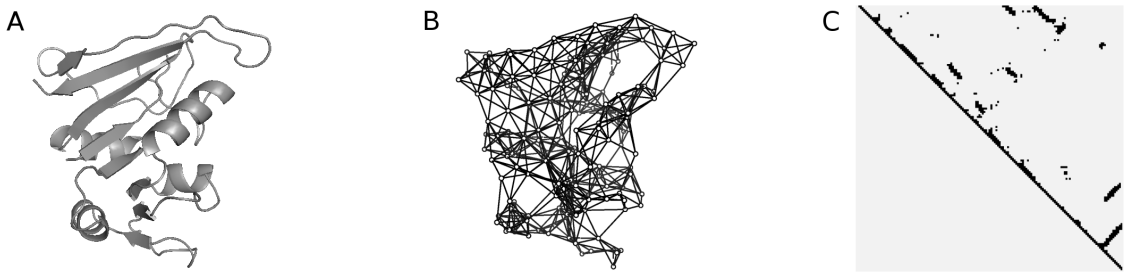

Figure 1.5: Graph representation of protein structures. (A) Cartoon representation of a protein. (B) Contact graph representation. (C) Contact map representation.

distances between C-beta atoms, or between any two atoms of the two residues (Huan et al., 2005). Contacts have also been defined based on the delauny tesselation (de Berg et al., 2008), in which case no distance threshold needs to be defined (Taylor and Vaisman, 2006).

An alternative, but equivalent representation of a contact graph is a *contact map* (Figure 1.5 C). Formally, a contact map is a graphical depiction of the adjacency matrix of the contact graph. Contact maps have been used since the 1970s to depict contact patterns Phillips (1970) and to display contact formation in folding studies (Levitt and Warshel, 1975). Figure 1.5 shows a protein in cartoon representation, and the corresponding contact graph and contact map. The rows and columns of the contact map correspond to positions in the primary sequence. A dot in row i and column j represents a contact between residues number i and j. This planar representation gives a good overview over certain structural features. For example, alpha-helices can be seen as thick areas along the main diagonal. Likewise, parallel and anti-parallel beta strands can be seen as stretches of contacts which are parallel or orthogonal to the main diagonal, respectively. Because contact maps and contact graphs are just alternative representations of the same concept, we will use the words interchangeably in cases where the distinction is not relavant to the argument.

A simple way to convert from atomic coordinates to a contact graph is to measure the distances between any two C-alpha atoms and then apply the cutoff, e.g. 8Å. Not quite so simple, but nevertheless possible, is to convert from a contact graph back to atomic coordinates. Several methods for this procedure called *3D reconstruction* have been proposed. *Distance geometry* methods have been developed in the context of determining molecular structures from Nuclear Magnetic Resonance Spectroscopy (NMR) (Crippen and Havel, 1988). Recently, stochastic methods have gained popularity because they are faster and deliver results of simular quality (Vendruscolo et al., 1997). By making use of the contacts as distance constraints and additional physicochemical restraints, such as typical bond lengths and bond angles, it is possible to reconstruct a protein structure from its contact graph up to an average accuracy of 2Å C-alpha RMSD (Duarte et al., 2010). This value is close to the resolution of most experimental methods for structure determination. The deviation stems from the fact that the same contact graph represents an ensemble of structures which all satisfy the given set of distance constraints. Just like for NMR ensembles, this can be used to represent some degree of flexibility in the structure. Furthermore, the graph represen-

tation makes it possible to apply algorithms from graph theory to address common protein analysis problems.

1.2.3 Graph-based methods in protein structure analysis

Graph-based models have been successfully applied to various problems in computational structural biology. The problem in *domain decomposition* is to devise an automatic method which divides a multi-domain protein into its subunits. Xu and colleagues showed that when treating the protein as a flow network, the domain boundaries coincide with the bottlenecks or *minimum cuts* in the network (Xu et al., 2000). These can be found by the Ford-Fulkerson algorithm known from graph theory (Cormen et al., 2001). Network models have also been used to predict properties of individual residues in proteins. Several studies showed that network centrality of residues which are solvent accessible correlates with functional importance (Amitai et al., 2004; del Sol et al., 2005). Paszkiewicz et al. used local graph measures to predict viable circular permutation sites which can be used to design split enzyme reporter proteins (Paszkiewicz et al., 2006). A third class of problems where graph models have been successfully used is protein structure alignment (Caprara et al., 2004). This topic will be discussed in detail in the following chapter.

Chapter 2

Graph-based multiple structure alignment

2.1 Introduction

An alignment is a hypothesis about the evolutionary relationship between two biological sequences. The task to find an "optimal" alignment in all its subforms is so prevalent in modern molecular biology that the original publication describing the *Blast* heuristic alignment tool (Altschul et al., 1990) is among to most cited papers in all of the life sciences (Russo and Bunk, 1999). Alignment algorithms can be subdivided into sequence- and structure alignment methods. In both cases, the goal is to find a mutual mapping between the positions in the sequences that best describes their evolutionary relationship. The two classes of methods differ in what information is being used to infer the alignment and in the algorithmic approaches. A second possible classification is to distinguish between pairwise and multiple alignment methods. Quite simply, in the pairwise case we are dealing with two sequences while the multiple case is a generalization to three or more sequences.

In this chapter, we investigate the problem of *multiple structure alignment* (MStA). According to the classification outlined above, this can be considered the most challenging variety. Even in the pairwise case, structure alignment is an NP-hard problem (Goldman et al., 1999) which implies that it can not be solved exactly for non-trivial instances of the problem. Caprara et al. have stated that the theory to treat structure comparison is *"almost nonexistent, as the problems are a blend of continuous-geometric and combinatorial-discrete mathematics"* (Caprara et al., 2004).

What adds to the complication is that there is no general agreement on how to score different alignments and how to characterize the optimal one. The authors of different methods have come up with different criteria for on optimal multiple alignment and hence the algorithms solve slightly different variations of the problem (Konagurthu et al., 2006; Shatsky et al., 2004).

For pairwise structure alignment, *contact map overlap* is an established framework (see next section for a review of other approaches). A thorough analysis of contact

map alignment for the multiple case is missing.

Here, we show how contact map alignment can be generalized to the multiple case using the sum of scores as the objective function to be maximized. Using this definition, we find a surprising connection to a seemingly unrelated problem: the theory of the graph sample mean recently introduced by Jain and Obermayer (Jain and Obermayer, 2008). We prove that the problem to find an optimal MStA is equivalent to the problem of finding a sample mean of graphs. This theoretical results has the following important implications:

1. Methods which have been developed for calculating the sample mean can be applied to MStA.
2. The results for the sample mean give a theoretical justification for iterative MStA methods such as the one presented in section 2.2.3.
3. The sample mean definition opens the door for further machine learning methods to be applied to proteins modelled as graphs.

In accordance with 1 and 2 we propose a new MStA method which approximates the optimal multiple alignment. We show that this method has some desirable properties and benchmark it against other recent multiple structure alignment methods.

2.1.1 Related work

Several structure alignment methods have previously been proposed which can be broadly categorized into three classes:

(1) Methods which reduce the problem to one dimension by considering local structural environments around the residues and then apply methods similar to those used for sequence alignment (Karpen et al., 1989; O'Sullivan et al., 2004; Tyagi et al., 2006; Schenk et al., 2008; Margraf et al., 2009). These methods are generally very fast but have the disadvantage that they often have problems with large insertions or deletions between diverged structures. Methods which are based on aligning secondary structure elements such as SSM (Krissinel and Henrick, 2004) can also be considered a subclass of this approach even though some have also been generalized to allow for non-sequential alignments and hence use methods which are quite different from classic dynamic programming (Guerler and Knapp, 2010).

(2) Algorithms which attempt to find maximal common substructures by rigid-body coordinate transformations (Sutcliffe et al., 1987; Russell and Barton, 1992; Gerstein and Levitt, 1996; Menke et al., 2008; Ilinkin et al., 2010). These methods work well for protein families with well conserved cores but fail for cases of conformational changes or relative domain reorientation.

(3) Methods which align distance- or contact matrices either of fragments (Holm and Sander, 1993; Shindyalov and Bourne, 1998) or simultaneously for the whole structure (Caprara and Lancia, 2002; Caprara et al., 2004; Xie and Sahinidis, 2007). The main motivation for these methods is the observation that the residue-contact pattern is the most deeply conserved feature in distantly related proteins (Lesk and Chothia, 1980).

Most structure alignment algorithms can only be applied to pairs of structures. Different approaches exist to generalize structure alignment to the multiple case, resembling the ones used for multiple sequence alignments. One common approach is to first perform all-against-all pairwise comparisons and then progressively build up a multiple alignment along a guide tree which is derived from the pairwise similarities with a method such as neighbor joining (Sali and Blundell, 1990; May and Johnson, 1995; Lupyan et al., 2005; Konagurthu et al., 2006). The methods by Ye (Ye and Janardan, 2004) and by Ilinkin (Ilinkin et al., 2010) use a similar, iterative approach but attempt to minimize the distances to a consensus structure which is build along way. Some methods first identify local, pairwise similarity patterns and then assemble them into a multiple alignment using a combinatorial or heuristic procedure (Guda et al., 2001; Menke et al., 2008). A third approach is to directly optimize a scoring function in the space of multiple alignments, for example using stochastic sampling methods (Godzik and Skolnick, 1994).

2.1.2 The sample mean of graphs

Many objects we deal with in bioinformatics can naturally be represented as graphs. In contrast, most algorithmic tools in machine learning and statistics are only defined for numbers or vectors. To generalize such methods and make them potentially applicable to graphs, we need concepts equivalent to the common operations defined for vectors. One such concept is the sample mean. Many useful algorithms such as *k-means clustering* rely on a suitable definition of the sample mean. If we can define such a concept for graphs, we can adapt the respective algorithms to be applicable to graph data.

A theory for the sample mean of graphs has recently been introduced by Jain & Obermayer (Jain and Obermayer, 2008). In section 2.2.2 we will show how this theory relates to protein structure alignment.

2.2 Results

2.2.1 A rigorous definition of the MStA problem

Contact Map Alignment is an established framework for pairwise protein structure comparison (Caprara et al., 2004). The objective is to find an alignment that maximizes the number of shared contacts in the two proteins to be aligned.

Here, we generalize the definition to the multiple case by considering the sum of pairwise scores as the objective function to be maximized. To do this in a formal way, we have to first establish some definitions:

Alignment of Weighted Graphs

A *weighted graph* is a triple $X = (\mathcal{V}, \mathcal{E}, \omega)$ consisting of an ordered set $\mathcal{V} = \{1, \ldots, n\}$ of vertices, a set $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ of *edges*, and a weight function

$$\omega : \mathcal{V} \times \mathcal{V} \to \mathbb{R}, \quad (i, j) \mapsto x_{i,j}$$

such that edges have positive weights, non-edges have weight zero, and vertices may have any value as weight. The vertex weights of X induce a partition of the set $\mathcal{V}$ into two disjoint subsets

$$\mathcal{R} = \{i \in \mathcal{V} \,|x_{ii} \neq 0\} \quad \textit{and} \quad \mathcal{G} = \{i \in \mathcal{V} \,|x_{ii} = 0\}$$

The elements of $\mathcal{R}$ are the *residues* of X and the elements of $\mathcal{G}$ are its *gaps*. The number of vertices of a weighted graph X is its *order*, written as $|X|$. We identify a vertex $i \in \mathcal{V}$ with its *position* i in the sequence $1, \ldots, |X|$.
The set of vertices of S is also referred to as $\mathcal{V}(X)$, its set of residues as $\mathcal{R}(X)$, and its set of gaps as $\mathcal{G}(X)$. Similarly, by $\mathcal{E}(X)$ we denote the set of edges of X. A graph X is completely specified by its *matrix representation* $\boldsymbol{X} = (x_{ij})$ with entries $x_{ij} = \omega(i, j)$.
Let $X = (\mathcal{V}, \mathcal{E}, \omega)$ and $X^\alpha = (\mathcal{V}^\alpha, \mathcal{E}^\alpha, \omega^\alpha)$ be weighted graphs. We call X^α an *alignment* of X if there exists a bijection $\alpha : \mathcal{R}(X) \mapsto \mathcal{R}(X^\alpha)$ such that

1) $i < j \Rightarrow \alpha(i) < \alpha(j)$ (order preserving)

2) (i, j) (2) $\mathcal{E} \Leftrightarrow (\alpha(i), \alpha(j)) \in \mathcal{E}^\alpha$ (structure preserving)

for all $i, j \in \mathcal{R}(X)$. By $\mathcal{A}(X)$ we denote the set of all finite alignments of $X \in \mathcal{X}$.
A *contact graph* is a weighted graph $X = (\mathcal{V}, \mathcal{E}, \omega)$ with

1) $\omega(i, j) \in \{0, 1\}$ for all $i, j \in \mathcal{V}$

2) $(i, j) \in \mathcal{E} \Rightarrow (j, i) \in \mathcal{E}$

3) $(i, j) \in \mathcal{E} \Rightarrow \omega(i, i) = \omega(j, j) = 1$

A *contact map* of a contact graph is its matrix representation.

Pairwise Alignment

Let $X, Y \in \mathcal{X}$ be graphs. A *pairwise alignment* of X amd Y is a pair $(X^\alpha, Y^\alpha) \in \mathcal{A}(X) \times \mathcal{A}(Y)$ of alignments such that

1) $\mathcal{V}(X^\alpha = \mathcal{V}(Y^\alpha))$

2) $\mathcal{G}(X^\alpha) \cap \mathcal{G}(Y^\alpha) = \emptyset$

The first condition states that the aligned graphs have the same number of vertices. From the second condition follows that there is at least one residue at each position. Both conditions bound the number of vertices by

$$|\mathcal{V}(X^\alpha)| = |\mathcal{V}(Y^\alpha)| \leq |X| + |Y|.$$

Let $\mathcal{A}(X,Y) \supset \mathcal{A}(X) \times \mathcal{A}(Y)$ denote the set of all pairwise alignments of X and Y. We measure the quality of an alignment by the *score function*

$$f : \mathcal{A}(X,Y) \to \mathbb{R}, \quad (X^\alpha, Y^\alpha) \mapsto \frac{1}{2} \sum_{i \in \mathcal{V}} \sum_{j \in \mathcal{V}} x_{ij}^\alpha \cdot y_{ij}^\alpha \tag{2.1}$$

where $\mathcal{V}$ is the common vertex set of the alignment and $\boldsymbol{X^\alpha} = (x_{ij})^\alpha$ and $\boldsymbol{Y^\alpha} = (y_{ij}^\alpha)$ are the matrix representations of X^α and Y^α. An *optimal pairwise alignment* is a pairwise alignment with maximal score.

Multiple Alignment

Let $S = \{X_1, \ldots, X_k\} \subseteq \mathcal{X}$ be a set of k graphs. A *multiple alignemnt* of S is a k-tuple $S^\alpha = (X_1^\alpha, \ldots, X_k^\alpha) \in \mathcal{A}(X_1) \times \cdots \times \mathcal{A}(X_k)$ such that

1) $\mathcal{V}(X_1^\alpha) = \cdots = \mathcal{V}(X_k^\alpha)$
2) $\mathcal{G}(X_1^\alpha) \cap \cdots \cap \mathcal{G}(X_k^\alpha) = \emptyset$

By $\mathcal{A}(\mathcal{S})$ we denote the set of all multiple alignments of $\mathcal{S}$. We measure the quality of a multiple alignment of $\mathcal{S}$ by the score function

$$F : \mathcal{A}(\mathcal{S}) \Rightarrow \mathbb{R}, \quad \mathcal{S}^\alpha \mapsto \sum_{p=1}^{k} \sum_{q=p+1}^{k} f(X_p^\alpha, X_q^\alpha) \tag{2.2}$$

where f is the score function of pairwise alignment as defined in Equation (2.1) and $\mathcal{S}^\alpha = (X_1^\alpha, \ldots, X_k^\alpha)$. We call $F(\mathcal{S}^\alpha)$ the *Gram sum* of $\mathcal{S}^\alpha \in \mathcal{A}(\mathcal{S})$. An *optimal multiple alignment* is a multiple alignment with maximal Gram sum.

This gives us a formal definition of an optimal alignment. The next question is how to find such an alignment. In general, the problem of maximizing the score in equation 2.2 is $\mathcal{NP}$-hard because it comprises the contact map overlap problem as a special case (Goldman et al., 1999), so we have to rely on heuristic methods (Garey and Johnson, 1979). Before we introduce such a heuristic method in section 2.2.3, we first show the links between multiple structure alignment according to the definition we have just introduced and the sample mean of graphs.

2.2.2 Equivalence of MStA and the sample mean of graphs

This section shows that MStA is equivalent to the problem of determining a sample mean of a set of graphs. The proofs for the results presented here are given in Appendix A.

The optimization formulation of the sample mean for vectors $x_1, \ldots, x_k \in \mathbb{R}^n$ minimizes the cost function

$$J_{vec} : \mathbb{R}^n \mapsto \mathbb{R}, \quad \boldsymbol{x} \mapsto \frac{1}{2} \sum_{p=1}^{k} d(\boldsymbol{x}, \boldsymbol{x_p})^2$$

where $d(\boldsymbol{x}, \boldsymbol{y}) = \|\boldsymbol{x} - \boldsymbol{y}\|$ is the Euclidean distance on $\mathbb{R}^n$. The cost function J_{vec} is smooth and convex. The solution of the above optimization problem can be given in closed form and is the standard formula of the sample mean

$$\boldsymbol{m} = \frac{1}{k} \sum_{p=1}^{k} \boldsymbol{x_p}$$

Since it is unclear how to define an addition on contact graphs, we resort to the optimization based formulation of the sample mean and adopt it to a sample mean formulation of contact graphs. Suppose that $\mathcal{S} = \{X_1, \ldots, X_k\}$ are contact graphs. In analogy to the standard sample mean, we set up the cost function

$$J : \mathcal{X} \to \mathbb{R}, \quad X \mapsto \frac{1}{2} \sum_{p=1}^{k} D(X, X_p)^2 \tag{2.3}$$

where D is a distance metric on $\mathcal{X}$ derived from the score function f of the pairwise contact map overlap problem (see Equation (2.1)) in the same way as the Euclidean metric d is derived from the inner product on $\mathbb{R}^n$. A *structural sample mean* of $\mathcal{S}$ is a weighted graph X with minimal cost $J(X)$.
Suppose that $\mathcal{S} = \{X_1, \ldots, X_k\}$ is a set of contact graphs. The following statements hold:

Result 1: The cost function J given in Equation (2.3) is locally Lipschitz and has a global minimum.

Result 2: Let $M \in \mathcal{X}$ be a structural sample mean of $\mathcal{S}$. The matrix representation $\boldsymbol{M}$ of M is the standard sample mean of the martrix representations of an optimal multiple alignment $\mathcal{S}^\alpha = \{X_1^\alpha, \ldots, X_k^\alpha\}$ of $\mathcal{S}$, i.e. M is of the form

$$\boldsymbol{M} = \frac{1}{k} \sum_{p=1}^{k} \boldsymbol{X}_p^\alpha$$

Result 3: Conversely, the standard sample mean of the matrix representations of an optimal alignment is a matrix representation of a structural sample mean.
The implications of the three statements are as follows: The first part of Result 1 gives rise to a subgradient method for determining a sample mean as described in the next section. The second part of Result 1 is in accordance with the fact that an optimal multiple alignment is not unique in general. From Results 2 and 3 follows that the problems of determining a structural sample mean and an optimal multiple alignment are equivalent. Hence, we can minimize Equation (2.3) to obtain a solution to the multiple alignment problem.

Now that we have established the equivalence between the sample mean of graphs and the multiple alignment problem, we can apply methods that have orignally been developed for finding the sample mean to finding alignments. Some of these methods are summarized in a recent publication by Jain et al. (Jain and Obermayer, 2009a). Here, we have adapted one such method to be applied to MStA.

2.2.3 A heuristic algorithm for the MStA problem

By Result 1, the cost function J of Equation (2.3) is locally Lipschitz. To minimize locally Lipschitz functions, the field of nonsmooth optimization offers a number of techniques (Mäkelä and Neittaanmäki, 1992). The simplest and probably the most used method for non-differentiable optimization are subgradient methods. The basic idea is to replace gradients by subgradients in classical steepest descent methods.

Algorithm 1 summarizes our method SMEG-Align, a subgradient method for MStA.

Algorithm 1 : SMEG-Align

Input: set $\mathcal{S} = \{X_1, \ldots, X_k\}$ of contact graphs
choose starting point $M_i \in \mathcal{X}$ and set $t := 0$
repeat
  set $M_{t,1} := M_1$
  for $i = 1, \ldots, k$ **do**
    choose $\left(M_{t,1}^{\alpha}, X_i^{\alpha}\right) \in \mathcal{A}(M_{t,1}, X_i)$
    determine step size $\eta_{t,i} > 0$
    set $M_{t,i+1} := M_{t,i}^{\alpha} + \eta_{t,i} X_i^{\alpha}$
  end for
  if $J(M_t) > J(M_{t,k+1})$ **then**
    set $M_{t+1} := M_{t,k+1}$
  end if
  set $t := t + 1$
until the maximum number of steps has been reached

Suppose that $(\boldsymbol{X}^{\alpha}, \boldsymbol{X}_p^{\alpha})$ are the matrix representations of an optimal alignment of the weighted graphs X and X_p. Then $\boldsymbol{X}^{\alpha} - \boldsymbol{X}_p^{\alpha}$ is a matrix representation of a subgradient of the p-th term $D(X, X_p)^2$ of $J(X)$ at X. Thus, to determine the subgradients, we need to solve a weighted version of the pairwise contact map overlap problem in each step. For solving the pairwise problem we use the *Bimal* algorithm proposed by Jain et al. (Jain and Obermayer, 2009b) because it can deal with weighted contact maps and provides a good tradeoff between computational speed and solution quality.

The final algorithm has been implemented in Java and can run on any platform with support for the Java Virtual Machine. In the following section we compare SMEG-Align against other current structure alignment algorithms.

2.3 Benchmarking

2.3.1 Experimental setup

We compared our algorithm *SMEG-Align* to three recent multiple structure alignment algorithms: *Paul*, the only other method we are aware of, which, like SMEG-Align, computes multiple alignments of contact maps (Wohlers et al., 2009); *Mustang*, a

well-known classical multiple structure alignment method based on rigid body superpositions (Konagurthu et al., 2006), and *Matt*, a recent method which has been particularly designed to allow for flexibility during the alignment (Menke et al., 2008).

We only included algorithms for which we could obtain a Linux executable so that we could perform the calculations under controlled conditions. The tests were performed on a Linux machine with a dual core AMD 64 X2 4000+ CPU and 4GB of RAM. All algorithms were run with default parameters.

2.3.2 Dataset

As a testset we used the Homstrad database which is the most widely used benchmark set for comparing multiple structure alignment methods (Mizuguchi et al., 1998). Homstrad contains manually curated alignments for protein families which span a variety of structural classes, protein sizes and degrees of divergence.

As we are interested in multiple alignments, we only considered families with at least three members. Some additional families were excluded for technical reasons, namely sequences which span multiple chains and very large families which caused memory problems on our test machine. The final testset with details about each family is listed in Table 2.1.

Each of the four algorithms was run once for each family. The resulting multiple alignments were then evaluated with the measures explained below.

2.3.3 Evaluation

We evaluated the differents alignments with respect to two different measures: the sum of shared contacts (SoSC) and the core size/core RMSD (nCore/cRMSD). We also compared the runtimes of the algorithms.

nCore/cRMSD

The most common way of comparing multiple structure alignments is by means of the number of residues in the common core (nCore) along with the core RMSD (cRMSD). The underlying assumption is that in a family of related structures, there is a common core of residues which is highly conserved. The quality of the superposition for a given core is measured by the *root mean square deviation* or *RMSD* averaged over the pairs of structures being aligned:

$$cRMSD = \frac{1}{N(N-1)} \sum_{i<j} RMSD_{i,j}, \tag{2.4}$$

where N is the number of structures and $RMSD_{i,j}$ is the usual pairwise RMSD calculated for the aligned core residues in structures i and j (Kabsch, 1976).

A good alignment algorithm should find the largest core which can be superimposed with the minimal cRMSD. These two measures are negatively correlated, such that there is a tradeoff between finding a larger core and aligning it with a smaller RMSD. The RMSD cutoff is often chosen depending on the application and differs between algorithms Kolodny et al. (2005).

A typical statement found in papers about MStA methods is of the form 'our algorithm identified a common core of size 25 with a core RMSD of 2.5Å'. It is not clear whether this is better or worse than a core size of 30 with an RMSD of 3.0Å. This makes it difficult to compare the results reported in different papers.

The second major shortcoming of the nCore/cRMSD measure is that it depends on a rigid-body superposition to calculate the RMSD and hence does not show how well the algorithms deal with conformational flexiblity. Instead, it measures how well the largest common cores can be superimposed.

Despite these disadvantages, the nCore/cRMSD is the most widely used measure to compare multiple structure alignments.

Sum of shared contacts (SoSC)

As an alternative to the above measure, we also evaluate the sum of shared contacts. This is the same score as given in equation 2.2. To evaluate this measure, the structures are converted to contact maps with a cutoff of 8Å between $C\alpha$ atoms. Then the number of shared contacts is evaluated pairwise and summed over all pairs of structures in the given alignment.

This value has the advantage that it measures the structural fit along the whole sequence as opposed to just considering a single conserved core. Because all regions in the structure contribute to the final score, it implicitly takes into account how well the algorithms deal with structural flexibility. Another advantage is that, as opposed to the nCore/cRMSD measure, it is a single score which can easily be compared across algorithms.

What needs to be kept in mind when interpreting the results for this scoring scheme is that two of the algorithms (SMEG-Align and Paul) explicitly strive to optimize the SoSC while other methods may be optimized for different evaluation schemes.

2.3.4 Results

The results of the benchmarking runs are shown in Figures 2.1, 2.2 and 2.3.

Sum of shared contacts

Figure 2.1 shows the performance of the different algorithms with respect to the SoSC score. The higher the curve, the better the performance of the algorithm. Paul performs best overall and finds the alignment with the highest score for 102 out of 125

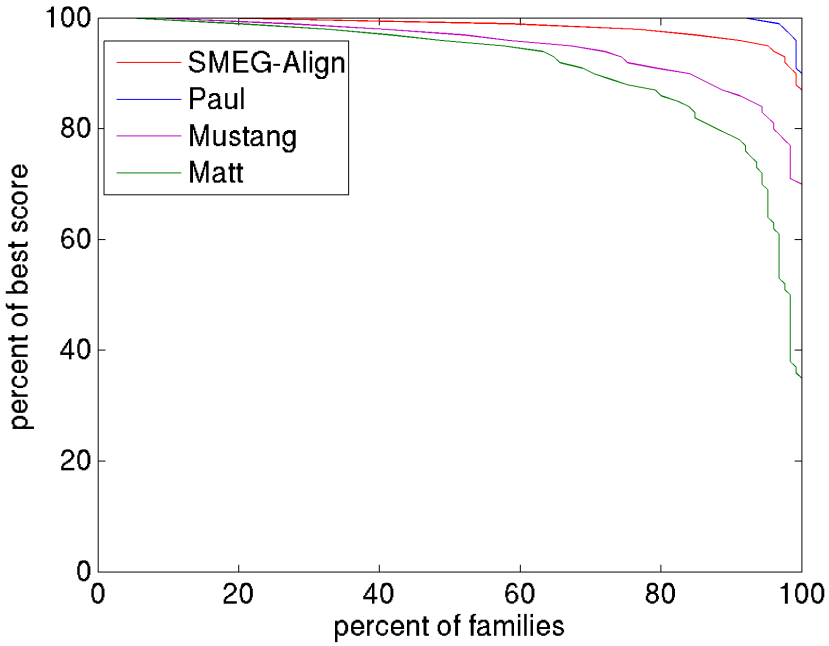

Figure 2.1: Comparison of multiple structure alignment algorithms. Alignment quality is measured as the number of shared contacts summed over all pairs of proteins in the family. The plot shows for the different algorithms what percentage of families x can be aligned within y percent of the score of the best performing algorithm. Higher scores indicate better alignments. Paul has the best overall performance closely followed by SMEG-Align.

families (curve coincides with the upper border of the plot). Its score is never below 90% of the best score. SMEG-Align closely follows with 120 out of 125 families within 95% of the best score and only one case where the score is worse than 90% of the best score. SMEG-Align performs better than any of the other methods in 10 out of 125 cases.

Runtime

A runtime comparison is shown in Figure 2.2. A data point (x, y) means that for x percent of the families, the respective algorithm was y times slower than the fastest algorithm. The blue curve shows that Paul always takes the longest time and can be more than 100 times slower than the fastest method for some families. Mustang and Matt are the fastest methods with very similar runtime for about 75% of the test families. SMEG-Align's performance (red curve) lies in the middle between Paul and Matt/Mustang. The general trend is that the more accurate algorithms take longer time to calculate the results.

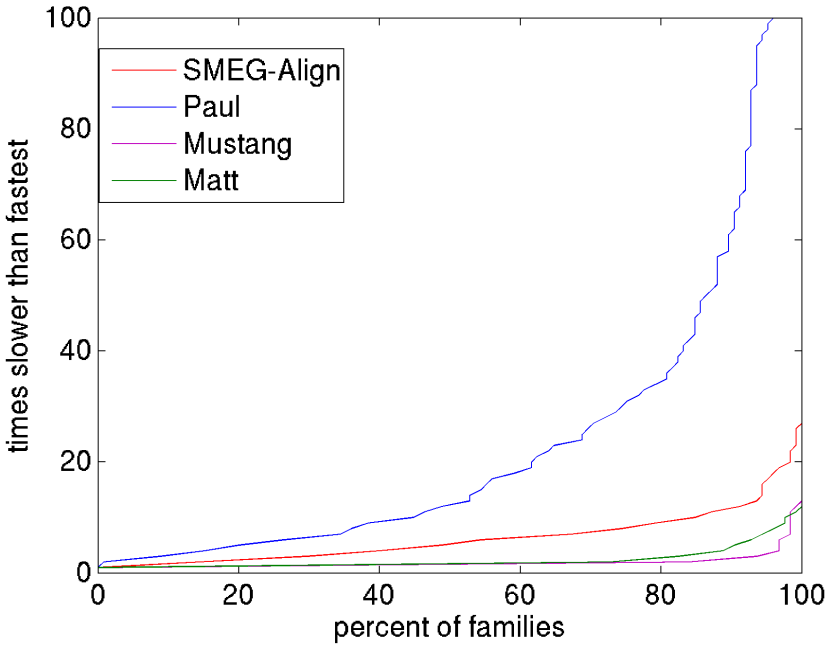

Figure 2.2: Comparison of runtimes. The plot shows for the different algorithms for how many percent x of the families the algorithm finishes within y times of the runtime of the fastest algorithm. A lower curve indicates a faster algorithm. The general trend is that the algorithms which produce better alignments also have longer runtimes.

nCore/cRMSD

As a second scoring scheme, we evaluated the classical core size/core RMSD measure. In Figure 2.3 the average core size is plotted over the core RMSD. This means that for a given core RMSD cutoff, we show what core size can be aligned within this cutoff. This value is first evaluated for each family, and then the average over all families is reported. The plot shows that different algorithms have strengths in different RMSD regimes. Up to 3.5Å, Matt performs best (green curve) with the largest core sizes. Above 3.5Å, SMEG-Align and Paul identify larger cores with almost indistinguishable results (blue and red curves). Mustang performs similar to SMEG-Align and Paul in the area up to 2.5Å, but identifies slightly smaller cores for RMSDs above 2.5Å. This shows that the graph based methods (SMEG-Align and Paul) excell at aligning diverged regions, while Mustang and in particular Matt, are better at finding highly conserved cores.

2.4 Discussion

2.4.1 Benchmark results

The method most similar in methodology to SMEG-Align is Paul and the two methods show very similar characteristics in the benchmarks. In the SoSC score, where the

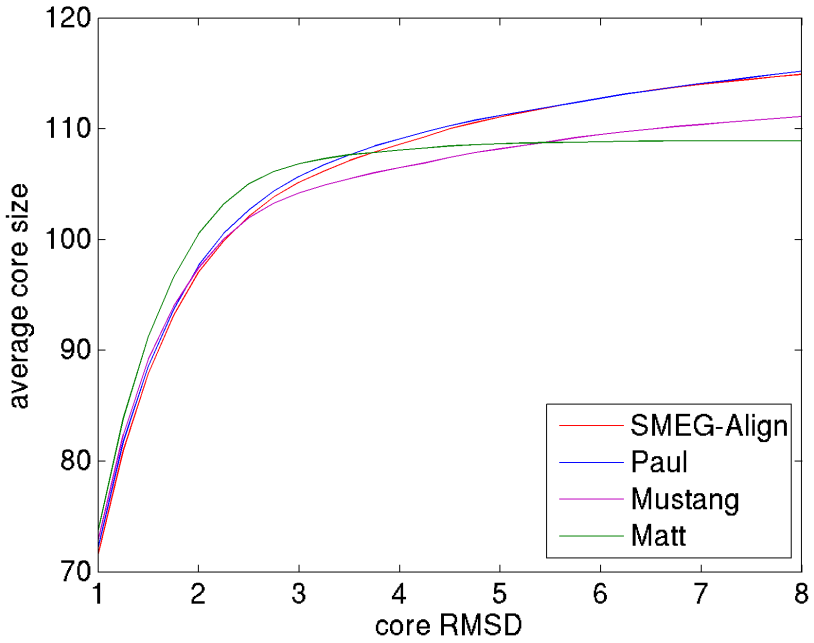

Figure 2.3: Comparison of multiple structure alignment algorithms with respect to core RMSD. The plot shows for a given core RMSD x, the size of the maximal core y (in number of residues) which can be superimposed with an average RMSD of at most x. The reported core size is averaged over the families from the Homstrad dataset. Larger core sizes indicate better alignments. Different algorithms have strengths in different RMSD regimes with Matt performing best for highly conserved cores with RMSDs up to 3.5 and SMEG-Align and Paul performing best for more diverse cores with RMSDs above 3.5.

two methods are directly comparable, Paul is more accuracte in most, but not all cases. Paul's overall better accuracy is paid for by an increased runtime. The two other methods are faster than both Paul and SMEG-Align but achieve lower scores, especially when considering regions outside highly conserved cores. In summary, graph based methods are the preferred choice when alignment accuracy along the whole sequence is important. Among the graph based methods, SMEG-Align achieves a very good compromise between accuracy and speed.

2.4.2 Advantages of graph-based alignment methods

When compared to other approaches, graph-based methods, such as SMEG-Align, have a number of specific advantages.

While classical methods based on rigid-body superpositions can only identify a single conserved core, graph-based methods can align all positions along the sequence. Flexibility is automatically taken into account because the optimization procedure automatically groups together regions in the structure, which have a similar contact pattern. The density of contacts within domains is higher than the one between domains. Thus, even if the relative orientiation of a domain compared to the rest of the

structure has changed during evolution, the contact pattern will be mostly conserved and the algorithm will group the similar domains together in the alignment.

Another interesting property is that the weighted sample mean contact map, which is being calculated on the way, is a representation of a consensus structure of the protein family being aligned. It shows which contacts are always present in representatives of the family and which are unique to a particular family member. In the next chapter on consensus contact prediction, we will make use of this property.

Moreover, the framework of graph alignment is independent of the particular context of protein structure alignment. The methods can easily be adapted to similar problems such as the alignment of other types of biological networks.

The main disadvantage of graph-based methods for practical applications is their increased runtime compared to alternative methods such as Matt or Mammoth.

2.5 Conclusion

In this chapter, we have shown how to generalize the alignment of contact maps to the multiple case. This allows to define MStA in a mathematically rigirous way based on maximizing the number of shared contacts. We have shown that in this form, MStA is equivalent to calculating the sample mean of graphs. The main practical implication of this result is that algorithms developed for finding the sample mean can now be applied to MStA. We have adapted one such method to the MStA problem and compared it to other existing methods. We have seen that our method performs well when compared to classical MStA methods. Compared to other graph based alignments, which provide similar advantages, our method shows an excellent tradeoff between solution quality and speed.

Table 2.1: Overview of Homstrad dataset

				SMEG-Align		Paul	
Family	Size	øLength	Class	Score	Time	Score	Time
AAA	4	314	alpha beta	5911	107.8	6216	639.0
aabp	3	232	alpha beta	3044	27.7	3045	107.0
ace	6	534	alpha beta	33769	363.6	34201	1312.2
ACPS	3	109	alpha plus beta	428	10.4	438	60.5
adh	5	373	multi domain	18883	126.3	18875	204.5
adk	9	207	alpha beta	27541	74.2	28040	263.9
AhpC-TSA	3	195	alpha beta	2209	19.5	2241	92.3
annexin	6	317	all alpha	21445	94.1	21432	101.6
apbact	3	280	alpha beta	3062	57.6	3281	478.7
Bcl-2	3	166	membrane bound all alpha	1589	12.6	1650	30.2
BIR	4	110	small	1852	9.4	1872	60.1
blmb	6	241	alpha plus beta	12758	72.0	13404	587.9

– continued on next page

Table 2.1 – continued from previous page

Family	Size	øLength	Class	SMEG-Align Score	SMEG-Align Time	Paul Score	Paul Time
C1	3	56	small	552	2.5	552	33.3
C2	3	130	all beta	1443	26.7	1529	309.6
cat	3	567	multi domain	6306	164.5	6337	680.1
cat3	4	232	alpha beta	5405	35.9	5538	105.6
CBD-3	3	153	all beta	1779	50.5	1864	339.1
cbm12	3	52	all beta	487	33.7	489	176.0
cbp	8	159	all alpha	15444	34.0	15724	88.4
ChtBD	5	42	small disulphide	193	2.8	193	22.8
cks	3	81	alpha plus beta	617	3.4	634	9.5
Colicin	3	194	membrane bound all alpha	2512	61.3	2614	420.2
COX2	3	212	all beta	2101	25.1	2117	74.2
COX3	3	239	membrane bound all alpha	2551	27.8	2595	94.0
csp	3	67	all beta	881	3.0	884	4.1
Cu-nir	3	334	all beta	4998	55.8	5002	45.8
CUB	3	110	all beta	1445	6.6	1445	27.1
cyclin	3	252	all alpha	2705	31.1	2742	119.8
cyclo	6	171	all beta	10941	35.2	11182	50.6
cyt3	6	110	all alpha	5504	11.2	5568	33.5
cyt5	6	81	all alpha	4574	7.5	4597	30.7
cyto	3	154	all alpha	1958	11.7	1977	9.7
cytprime	4	127	all alpha	3060	10.9	3069	28.5
DEATH	7	107	all alpha	6917	15.1	7130	77.0
dhfr	4	172	alpha beta	4057	19.1	4085	80.7
DHH	3	185	alpha beta	4325	45.2	4327	45.8
DHHA2	3	119	alpha beta	4325	47.1	4327	45.8
DISIN	3	62	small disulphide	623	2.9	624	27.3
DNA-PPF	4	114	alpha plus beta	1042	15.8	1042	25.6
dsrm	3	82	alpha plus beta	292	5.5	304	4.8
dutpase	3	124	all beta	1457	7.8	1460	35.1
EF-TS	3	141	alpha plus beta	530	17.3	553	66.8
EF1BD	3	90	alpha plus beta	1064	4.9	1067	9.8
EFTU-C	3	97	all beta	5549	83.4	5568	91.5
ENTH	3	225	all alpha	2193	24.3	2246	87.0
fer2	13	98	small	29429	26.6	30372	157.7
ferritin	4	166	all alpha	4251	17.0	4313	14.5
Filamin	3	104	all beta	429	10.0	429	12.5
GAF	3	156	alpha plus beta	636	24.0	698	163.5
GATase	3	209	alpha beta	2449	76.7	2219	656.9
GBP-PSP	3	23	small	202	0.8	205	22.4
ghf11	5	185	all beta	8329	35.7	8437	58.8
gluts	14	215	multi domain	79329	105.6	79395	291.1

– continued on next page

Table 2.1 – continued from previous page

Family	Size	øLength	Class	SMEG-Align Score	SMEG-Align Time	Paul Score	Paul Time
gpr	3	154	all beta	2130	13.1	2130	55.1
hexapep	3	160	all beta	2093	43.0	2389	553.0
hip	5	74	small	2705	5.8	2684	23.6
hla	5	178	alpha plus beta	12334	56.9	12333	44.4
HMA	3	70	alpha plus beta	908	3.4	908	8.9
hpr	5	87	alpha plus beta	3935	7.5	3932	13.1
igC1	5	97	all beta	6470	39.0	6474	45.9
il8	11	68	alpha plus beta	12685	8.9	12792	29.7
int	5	192	alpha beta	7595	33.2	7660	70.6
kunitz	10	57	small disulphide	10497	6.4	10513	20.4
LDLa	4	41	small disulphide	772	1.7	762	5.4
LIM	5	69	small	2414	6.7	2420	28.2
lipase	5	447	multi domain	21923	180.3	21945	217.5
Lipase-3	3	266	alpha beta	3623	36.7	3644	142.9
LMWPc	3	157	alpha beta	2184	13.3	2184	16.0
LRR	3	419	alpha beta	5276	105.6	5414	373.9
LuxS	4	149	alpha plus beta	3396	14.1	3445	20.0
lyase-1	5	444	all alpha	17725	180.8	17930	768.7
MCR-beta	3	252	all alpha	6693	104.7	6693	66.3
mdd	3	385	multi domain	5613	82.5	5618	59.5
MHC-II-C	8	99	all beta	20053	45.8	20412	78.4
MHC-II-N	13	83	alpha plus beta	56605	78.8	57627	178.3
MIF	3	115	alpha plus beta	1467	6.9	1468	19.2
mthina	3	31	small	322	1.1	322	6.8
mthinb	3	30	small	264	1.0	268	7.5
mycin	4	111	all beta	2786	9.0	2795	25.4
neurotox	3	34	small disulphide	375	1.2	375	3.5
OTCace	5	318	alpha beta	13911	97.0	14017	149.5
P	3	55	small	184	2.8	196	7.7
PAS	3	124	alpha plus beta	1382	7.9	1395	26.0
PDZ	6	93	all beta	5514	13.9	5678	31.8
pgk	4	405	alpha beta	11551	118.4	11562	137.7
phc	12	166	all alpha	45219	53.2	45444	120.0
phero	3	39	small disulphide	468	1.5	469	3.0
pilin	3	135	alpha plus beta	1134	10.2	1172	115.1
PK	4	490	multi domain	13796	173.6	13826	721.9
plantltp	5	88	all alpha	3211	7.2	3222	24.8
pnp	3	271	alpha beta	3175	36.7	3214	295.9
pp	3	36	small	338	1.1	340	2.8
profilin	5	128	alpha plus beta	5658	15.4	5708	39.9
protg	4	65	alpha plus beta	1371	3.6	1382	14.1

– continued on next page

Table 2.1 – continued from previous page

Family	Size	øLength	Class	SMEG-Align Score	SMEG-Align Time	Paul Score	Paul Time
prt	3	174	alpha beta	1678	15.4	1710	84.0
PSI-PsaE	3	69	all beta	763	2.9	767	5.8
PTB	3	147	all beta	1446	13.3	1503	111.8
ptpase	3	285	alpha beta	3694	37.9	3730	64.6
ricin	7	254	alpha plus beta	23219	83.1	23333	149.5
RING	3	58	small	596	3.7	606	21.9
rnase	3	104	alpha plus beta	1337	5.9	1339	7.3
rnh	3	141	alpha plus beta	1648	11.0	1666	48.1
RRF	4	184	alpha plus beta	4637	21.6	4641	20.6
rub	5	51	small	1908	2.5	1935	5.8
rvp	6	106	all beta	5734	11.6	6029	33.9
seatoxin	5	47	small disulphide	1649	2.9	1666	9.7
slectin	5	133	all beta	5585	15.9	5745	20.8
Sm	4	76	all beta	1789	4.7	1805	11.8
sodcu	7	152	all beta	14429	32.2	14787	81.3
svmp	3	199	alpha plus beta	2780	19.3	2780	16.3
thionin	3	46	small disulphide	594	1.5	594	5.1
thiored	6	95	alpha beta	4480	11.4	4664	61.4
TIG	6	84	all beta	50927	502.6	51038	983.0
TIL	4	66	small	1110	3.6	1222	20.0
Toprim	3	110	alpha beta	4041	123.9	3894	679.6
TPR	6	153	all alpha	9123	76.1	8931	225.1
trfl	7	526	alpha beta	42007	442.9	42200	935.5
UBQ	3	74	alpha plus beta	938	3.1	937	5.5
uce	13	149	alpha plus beta	45383	52.2	45724	154.8
UCR-TM	3	69	membrane bound all alpha	1196	9.0	1220	29.6
ung	3	224	alpha beta	2906	24.0	2908	101.7
UPF0076	3	125	alpha plus beta	1693	9.2	1699	8.0
xia	6	388	alpha beta barrel	26752	142.3	26781	166.8
YgbB	3	153	alpha plus beta	1997	12.9	2007	19.4

Chapter 3

Consensus prediction of inter-residue contacts

3.1 Introduction

After having discussed protein structure comparison in the previous chapter, we will now turn to the protein structure prediction problem. Here, we propose a new method for a specific subproblem of structure prediction, which is to predict intra-molecular contacts.

First, we will introduce the problem, review related work and discuss how consensus methods have been successfully used in bioinformatics. We will also briefly introduce how structure prediction methods are assessed in the Casp experiments. Then, we will introduce a new consensus method for residue-residue contact prediction which makes use of the sample mean as a key step in the prediction procedure. Finally, we compare the performance of the method to other state-of-the-art methods and discuss the results and implications.

3.1.1 Contact prediction

Contact prediction is the problem to determine, given only the protein sequence, which residues are in spatial proximity in the folded structure. This problem is closely related to structure prediction because knowing the exact distances of all residue pairs implies knowing the 3D structure. The reason for regarding contact prediction a seperate problem (for example in Casp) was that traditionally, very different algorithmic approaches were used for contact prediction than for predicting 3D coordinates directly.

So how can residue contacts be predicted from sequence? The earliest attempts used information about correlated mutations derived from multiple sequence alignments (Altschuh et al., 1987), an approach which has been very successful for RNA structures (Winker et al., 1990). This is based on the assumption that for two residues which are in physical contact, a mutation in one will often lead to a compensating

mutation in the other. This approach has been further refined by considering correlations between pairs of residues (Göbel et al., 1994) or even larger fragments around the positions of interest (Hamilton et al., 2004). Another simple approach is to derive contact propensity matrices from known protein structures, essentially estimating the pairwise interaction energies (Singer et al., 2002). But it has been argued that the information gained using this approach alone is largely due to hydrophobic effects (Cline et al., 2002). This observation, and the aim to reduce noise, led to the idea of using reduced alphabets (Pollock et al., 1999) or to replace the amino acids by vectors of physicochemical properties (Neher, 1994; Vicatos et al., 2005). Most current methods include information from the above approaches but combine them with other sequencial and structural features using a machine learning or statistical inference method such as neural networks (Lund et al., 1997; Fariselli et al., 2001; Pollastri and Baldi, 2002; Punta and Rost, 2005a; Vullo et al., 2006; Shackelford and Karplus, 2007; Tegge et al., 2009; Xue et al., 2009), support vector machines (Zhao and Karypis, 2003; Cheng and Baldi, 2007) or bayesian inference (Miller and Eisenberg, 2008). Features which have been identified as informative include sequence conservation (Fodor and Aldrich, 2004), residue separation in primary sequence (Olmea and Valencia, 1997), predicted secondary structure (Fariselli et al., 2001), solvent accessibility (Pollastri and Baldi, 2002), amino acid properties of the intermediate region between the two positions (Punta and Rost, 2005a), global amino acid composition (Cheng and Baldi, 2007) and contacts inferred from threading templates (Shao and Bystroff, 2003).

Despite some progress in recent years (Ezkurdia et al., 2009), the prediction accuracy remains low with 37% (see the references for an exact definition of this value) achieved by the latest methods (Vullo et al., 2006; Cheng and Baldi, 2007) which is not sufficient to infer protein structure from predicted contacts alone. Instead, contact information has been successfully incorporated into structure prediction pipelines where it is used to select good models among sets of pre-selected models or templates (Miller and Eisenberg, 2008; Latek and Kolinski, 2008; Tress and Valencia, 2010). It has also been used to estimate protein folding rates (Punta and Rost, 2005b) and for identifying unstructured regions (Schlessinger et al., 2007).

3.1.2 Consensus methods in bioinformatics

The term *consensus methods* subsumes approaches whose basic idea is to integrate several independent methods, which strive to solve the same problem, into a new and improved combined method.

Consensus methods have been successfully applied in several fields of structural bioinformatics such as secondary structure prediction (Cuff et al., 1998), domain assignment (Veretnik et al., 2004), template identification for homology modeling (Bujnicki et al., 2001) and transmembrane topology prediction (Klammer et al., 2009).

The principle layout of a consensus method is illustrated in Figure 3.1. The simplest such method will just reevaluate the inputs and select the one with the highest score as the new output. For the case of structure prediction, a method like this could for example be implemented as follows: Train a classifier, e.g. a neural network to select from the input models one that is estimated to have the highest quality. The features

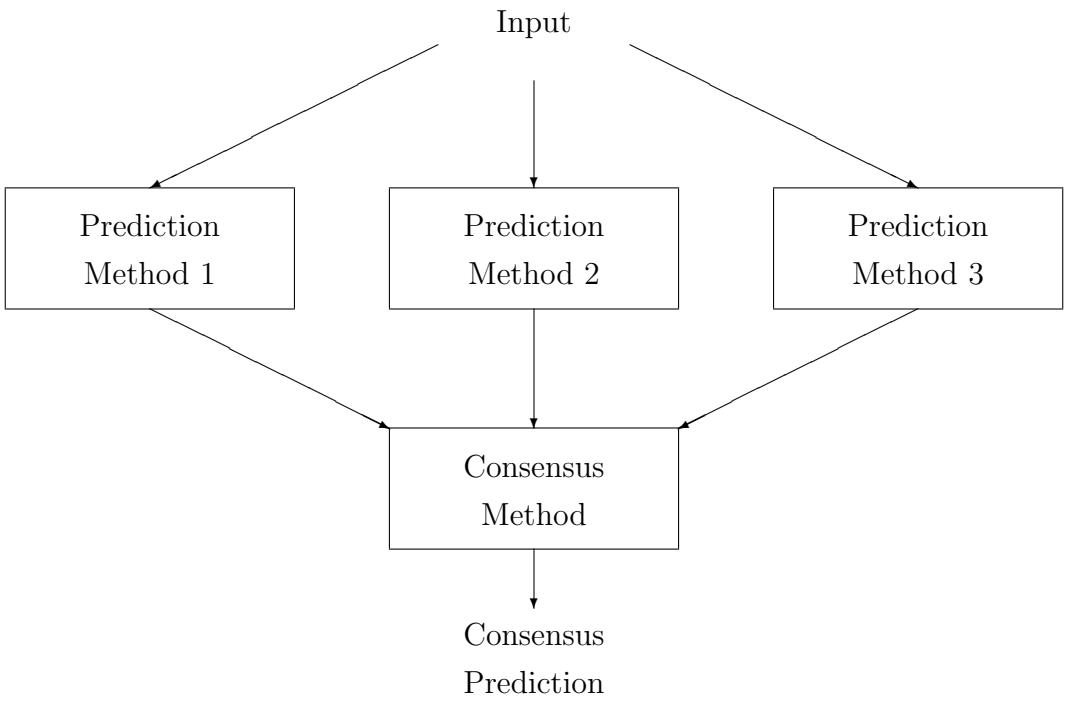

Figure 3.1: Outline of a consensus prediction method. The prediction results of several independent methods are combined by the consensus method into a new prediction.

for selection can be independent criteria such as energy terms, knowledge based scores or simply the identity of the input method. In the latter case the method would learn in which cases to trust one method more than another. In fact, many current consensus methods are based on the ideas outlined above (Bujnicki et al., 2001).

Whenever independent approaches exist, which at least partly use complementary information, there is a high chance that a consensus method can be found which improves the solution over the independent methods.

In the following sections we will explore the idea to build a consensus method for residue contact prediction.

3.1.3 The Casp experiment

The idea of the series of Casp experiments is to provide an independent assessment of structure prediction methods and to measure the progress of the field as a whole. The name Casp stands for Critical Assessment of Protein Structure Prediction Methods and the first Casp experiment was organized by John Moult in 1994 (Moult, 2005).

The basic idea is that protein structures which have recently been solved by X-ray crystallography or NMR, but which have not yet been published, are used as targets for a blind-test prediction experiment. The Casp organizers collect these unpublished structures directly from the crystallography or NMR labs and release the sequences on their website. Developers of prediction methods can sign up to the experiment and attempt to predict the structures for the targets. The predictions have to be sent in within a defined timeframe (usually about three weeks after the release date). They

are then evaluated by independent assessors who have access to the solved structures. The advantages of this set-up are that all participating methods are evaluated on the same test set, the test set is unknown to the method developers, so methods can not be over-optimized for the test set, and the evaluation is done independently and with the same criteria for all methods. This makes it possible to compare methods more objectively than by self-assessment by the authors of a particular method. Currently, the Casp experiment is conducted every two years during the summer months and the results are being presented at a subsequent conference in December of the same year. Apart from the main task, to predict protein 3D structures, several sub-categories of related prediction problems have been introduced and are assessed in a similar way:

- Secondary structure prediction
- Prediction of domain boundaries
- Residue-residue contact prediction
- Model quality prediction
- Identification of unstructured regions
- Prediction of ligand binding sites

It has been questioned whether the very results-oriented format of Casp is indeed the best way to promote the development of creative new methodologies but it is undisputed that it has pushed the progress in the field of structure prediction forward in the last decade (Moult, 2006). In fact, the success of the format has inspired several other assessment experiments such as *CAPRI* for protein-protein interactions (Janin et al., 2003) and *BioCreAtIvE* for information extraction and text mining (Hirschman et al., 2005).

3.2 Methods

The main question we want to address here is: Given the sequence of a target protein to be predicted, and a number of outputs from automatic and independent prediction methods, can we build a better model than the individual predictions?

In particular, we want to predict for a given protein, only knowing its sequence, which residues are in contact in the folded structure.

3.2.1 Source data

The first step is to collect as many independent predictions as possible, as inputs to our consensus method. For tertiary structure prediction, there are several online servers, which take the protein sequence as input and return a file in PDB format which contains the predicted structure. According to the Casp7 evaluation, the best 3D prediction methods are better at predicting contacts than the best dedicated contact prediction methods (Izarzugaza et al., 2007). In order to have the best quality of

input predictions, we use the 3D predictions converted to contact maps as inputs to our method.

The predictions made by the different automated prediction servers for previous Casp rounds can be downloaded from the Casp website. The predictions were originally generated by sending the target sequence to the servers which returned the predictions as PDB files. At the time of the prediction, the crystal structures were known neither to the servers nor to the developers.

As a benchmark set, we used the data from the last Casp8 round in 2008, where crystal structure have meanwhile been released, so that we could evaluate the results. The crystal structures were only used for benchmarking and not in any way for method development or prediction.

Altogether 76198 server models for the 128 prediction targets from Casp8 were downloaded (122 servers each contributing at most 5 predictions per target). 5 Targets were excluded because no crystal structure had been released by the time of the analysis. Individual server models were also excluded if they did not conform to the specified Casp format defined at `http://predictioncenter.org/casp8/index.cgi?page=format`.

3.2.2 Consensus contact prediction

Throughout this chapter, we use the following contact definition which is also used in the Casp experiments:

Two residues i and j are *in contact* if their $C\beta$ atoms ($C\alpha$ for glycine) are within 8Å of each other.

We use $C\beta$ instead of $C\alpha$ atoms because this retains information about the side-chain orientation of the individual residues. This has advantages when we later want to convert the predicted contacts maps to 3D structures (Duarte et al., 2010).

The basic steps of the consensus prediction method, which we call SMEG-CCP (for Sample MEan of Graphs Consensus Contact Predition) are as follows:

Input: The sequence and a number of structure predictions for the target protein

1. Convert the input structures to contact graphs
2. Calculate the sample mean of the input graphs (representing the ensemble of predictions)
3. Convert the weighted sample mean graph back to a binary contact map

Output: A contact prediction for the target protein

And for evaluation:

4. Compare the predicted contact map with the native contact map which has been obtained from the native 3D structure using the same procedure as in 1.

A graphical overview of these steps is given in Figure 3.2.

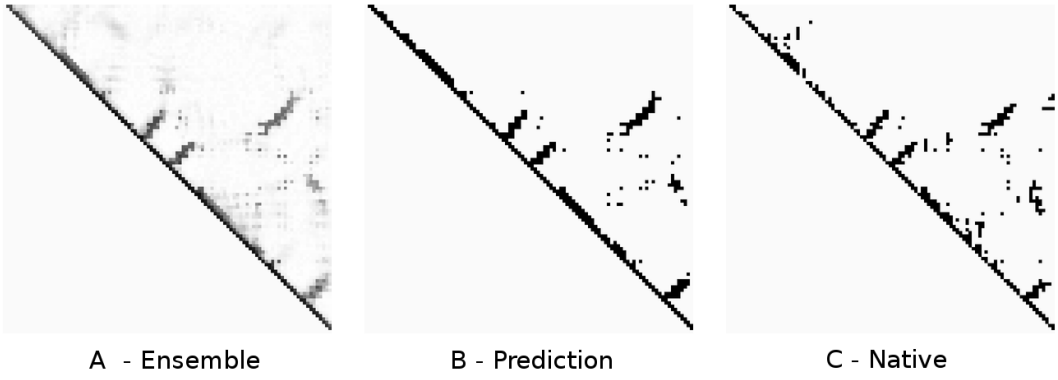

Figure 3.2: Principle of SMEG-CCP consensus contact prediction. Input structures are condensed into an ensemble by calculating the sample mean (A). The ensemble is then converted into a binary contact map (B), the prediction, e.g. by applying a cutoff. The prediction is evaluated against the native contact map (C).

The key step here is the calculation of the sample mean in (2). Since all input structures are predictions of the same protein, i.e. conformations of the same amino acid chain, mapping between the residues becomes trivial and no alignment is necessary. Then the calculation of the sample mean becomes simply a matter of edge counting. The set of nodes is the same as for all the predictions, namely the residues of the target protein. For every edge (i, j), the edge weight is the fraction of input structures, in which the edge is present. So for a protein of length n, where the N and C termini are in contact in half of the predicted models, the weight for edge $(1, n)$ would be 0.5. This number can also be interpreted as a propensity that the contact exists in the native structure. To finally decide which contacts to keep for the final model (which should be a binary contact map), it would be possible to apply a simple cutoff, e.g. 0.5. Instead of this 'majority vote' scheme, we found that an alternative approach gives more robust results across the whole range of target difficulties. We first estimate the number of contacts n we expect in the native structure, then rank the predicted contacts by the propensity and pick the top n from above. To estimate n, we again use a consensus argument. From a number of possible ways, the median number of contacts, over the set of input structures, worked best as a predictor. A simple interpretation is that the number of contacts is a rough measure for the packing density of a protein. Some of the input methods tend to overpack proteins while others tend to pack too loosely. By taking the median number of contacts we get rid of the outliers and obtain a good estimate of the native number of contacts.

3.2.3 Evaluation of predictions

We evaluated the predictions obtained by the consensus method against the native structures. The evaluation scheme and some concepts, which are used in the discussion of the results are explained below.

Comparing contact predictions - Accuracy and Coverage

A good prediction should predict as many of the native contact as possible. At the same time, it should not overpredict contacts that are not present in the native structure. With this in mind, we evaluated the predictions in terms of accuracy and coverage as defined below:

Acc = Contacts correctly predicted / Contacts predicted

Cov = Contacts correctly predicted / Native contacts

For a perfect prediction, both values should be 100%.

As usual in prediction, there is a tradeoff between accuracy and coverage. The relative importance of the two parameters depends on the particular application. We have shown in a previous study, that distance geometry reconstruction is quite tolerant to false negative contacts, but very sensitive to false positives (Sathyapriya et al., 2009). So in general, for predicting contact maps, it is important to keep the false positives rate low while maintaining a reasonable coverage.

Two other scores for evaluating contact predictions have been suggested: improvement over random and the *delta score* (see (Grana et al., 2005) for definitions), but they have been criticised as being too sensitive to random artifacts in lower quality predictions or, as Kevin Karplus put it, "*a way for people whose contact prediction methods don't work to try to salvage something from the failed effort*" (Personal communication).

Comparing 3D models - Global distance test

For measuring the similarity of two conformations of the same protein chain, for example, a prediction and a native structure, the GDT-TS score has been successfully used in the Casp experiments. It is better suited for evaluating 3D predictions than the classical root mean square deviation (RMSD) because it is more robust across a wide range of structural dissimilarity (Zemla, 2003). The score is defined as follows:

$$GDT - TS = 1/4(T_8 + T_4 + T_2 + T_1), \tag{3.1}$$

where T_n is the percentage of corresponding residues which are no more than n Å apart in the superposition which maximizes the T_n value. So the final GDT-TS value is based on four different superpositions.

Identical and near-identical conformations will have a GDT-TS score of 100.

Long range contacts

Contact predictions are commonly evaluated seperately for long range and short range contacts. The terms long range and short range here are not related to the distance of the two atoms being in contact but on their position in the sequence.

Following the definition used in Casp, we use a threshold of 24 and define long range contacts as follows:

For two residues i and j which are part of the same protein chain, the *sequence separation* is the absolute difference of the positions of i and j in the sequence. For simplicity, we use the same symbol for the residue and its position in the sequence and write $SeqSep(i, j) = |i - j|$.

A contact between residues i and j is called a *long range* contact if the sequence separation between i and j is greater than 24, i.e. $|i - j| > 24$. Otherwise it is called *short range*.

When evaluating our contact prediction results, we will show results seperately for all contacts, and for long range contacts. The reason for this is that short range contacts are inherently easier to predict, because they are dominated by secondary structure formation, for which reasonably accurate prediction alogorithms exist. We have shown however, that neither short-range nor long-range contacts alone are sufficient to determine the structure (Sathyapriya et al., 2009). So for practical applications, the performance for all contacts matters, while for measuring the methological progress, the results for long range contacts are more interesting.

3.3 Results and Discussion

3.3.1 Contact prediction results

Compared to input models

The minimal criterion a successful consensus method should fulfill is that it should perform better than the average or median of the input methods. To test this, we compared the accuracy and coverage of our predictions (measured against the native structure) to the respective accuracy and coverage of the input models. In the following, we combine accuracy and coverage into a single score by considering their average: (Acc+Cov)/2.

For all prediction targets, our predicted model is better than the median score. In fact, for half of the targets (61 out of 123), our model is better than any of the input structures. In 85% (104/123) of the cases, it is among the top 5 models (out of ≈100, depending on the particular target). Figure 3.3 shows the (Acc+Cov)/2 scores for one example target (T0409). The results for all targets are shown as boxplots in Figure 3.6.

When looking at long range contacts only, our model is ranked first in 35% (43/123) of the targets, among the top 5 in 67% (82/123) of cases and in the upper half for all but 5 of the targets (118/123). For all five targets, the average score of all models is below 17.0 so all models are essentially random. In this case the consensus does not improve the prediction.

So our method works slightly better for short range than for long range contacts.

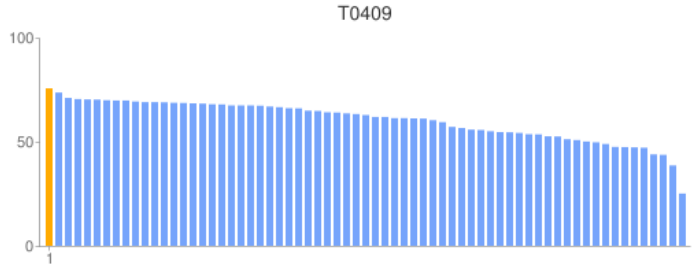

Figure 3.3: Comparison of model quality for all particicapting methods for Casp target T0409. Our method is shown in orange, other methods are shown in blue. The reported score is (Acc+Cov)/2. A higher bar indicates a better model.

Compared to other groups

The method's performance reported in the previous section compared our model to the input models disregarding the identity of the model's submitter. To compare our performance to those of other methods, we need to evaluate every groups's performance on all targets. We do this in terms of three scores, average score, median score and average rank, each taken over all targets a group submitted predictions for.

Some groups did not submit predictions for all targets. For a few targets this can be due to technical problems or a missed submission deadline. If too many targets were missed, the results can be biased towards easier targets. To make the results comparable, we included only groups with predictions for at least 100 out of the 123 targets.

Tables 3.1 and 3.2 show the top 10 methods ordered by median rank for all contacts (3.1) and for long range contacts (3.2). For both cases, our method performs better than any other method.

Table 3.1: Top prediction groups for Casp8 targets (all contacts)

Method	Average Score	Median Score	Average Rank	Median Rank
SMEG-CCP	79.4610	81.0545	3.5935	2
Zhang-Server_TS1	75.5711	77.6681	15.6829	12
HHpred5_TS1	73.7607	77.1055	21.7724	17
BAKER-ROBETTA_TS1	74.2328	75.7567	23.3577	19
MULTICOM-REFINE_TS1	73.8106	76.8107	22.0163	19
HHpred2_TS1	73.9589	77.7953	22.4146	19
MUProt_TS1	73.7155	76.4825	22.2927	20
RAPTOR_TS1	74.2043	76.3713	21.4309	20
Phyre_de_novo_TS1	73.7998	75.9413	23.0488	21
METATASSER_TS1	73.4942	75.7446	24.3496	21

Table 3.2: Top prediction groups for Casp8 targets (long range contacts)

Method	Average Score	Median Score	Average Rank	Median Rank
SMEG-CCP	62.8376	67.8022	6.2439	3
Zhang-Server_TS1	56.7753	63.1944	17.4065	14
HHpred5_TS1	54.7855	60.335	22.0569	16
Phyre_de_novo_TS1	54.2401	59.7522	23.8049	21
Pcons_multi_TS1	52.8802	58.8123	25.2195	22
HHpred4_TS1	54.6650	62.8341	24.6179	22
HHpred2_TS1	54.9869	63.3709	23.6992	23
GS-KudlatyPred_TS1	53.6903	60.6496	25.5537	23
MULTICOM-REFINE_TS1	54.0152	61.0651	24.8455	24
RAPTOR_TS1	54.7583	61.5955	23.6423	24

3.3.2 Independent Casp evaluation

We also submitted blind-test predictions generated with our method for the contact prediction category of the Casp9 experiment in 2010. The final assessment will not be available before late 2011 but a preliminary evaluation was shown at the Casp9 meeting in Asilomar, California in December 2010.

Table 3.3 has been reproduced from the results presented at the meeting.

The evaluation criteria differ slightly from the ones used for our own benchmark but are likely to be similar to the ones used in the Casp8 assessment (Ezkurdia et al., 2009). The final evaluation criteria will be published in a special issue of *Proteins* about Casp9 in late 2011.

According to the prelimiary results, our method performs best among the participating contact prediction methods. Interestingly, the method ranked second, with a very similar performance, is also a consensus method.

From contacts to 3D predictions

So far, we have evaluated our method at the contact level and saw that it performs better than other methods at predicting native contacts. The next question is whether we can use this knowledge about native contacts to also predict better 3D structures. To address this question, we generated 3D structures from our contact predictions using a distance geometry algorithm (Ponder and Richards, 1987a) and compared them to the input structures in terms of the GDT-TS score compared to the native (X-ray or NMR) structure. However, from the generated models, only 14% were above the average score of the input models, and only one model was in the top 20% of predictions. Visual inspection of the generated models showed that the overall fold was often similar to the best models, but the local structure such as secondary structure elements was consistently inaccurate. One possible reason for this loss of information from the best contact information to below-average 3D models is the distance geometry algorithm. Distance geometry was originally developed for NMR spectroscopy

Table 3.3: Top contact prediction groups in Casp9 - The evaluation criteria are explained in (Ezkurdia et al., 2009)

		L/5	
	Group Number	Nb Targets	Z, total
SMEG-CCP	391	27	2.3913
MULTICOM	490	27	2.388
MULTICOM-CLUSTER	2	25	1.2584
Infobiotics	51	28	1.0733
ProC_S3	138	24	1.0119
Distill	214	28	0.8809
SAM-T08-server	103	28	0.8400
MULTICOM-CONSTRUCT	80	26	0.7761
ProC_S1	375	25	0.7403
SAM-T06-server	244	25	0.6783
MULTICOM-REFINE	119	26	0.6739
PSICON	422	28	0.6277

where distance constraints are derived from atomic couplings in proteins. Since these constraints are derived from physical molecules, they are not expected to contain contradictory information. Such contradictory constraints can appear in the predicted contact maps, and may mislead the algorithm which attempts to fulfill them all, resulting in biophysically unfavorable geometries. Another reason is that the established methods for tertiary structure prediction make much use of known native structures in the form of fragments or templates. A reasonable choice of a template or fragment results in locally very accuracte geometries since they were derived from native structures. In the distance geometry procedure, no information from related structures is used such that local structure is often inaccurate despite an overall accurate fold.

A promising avenue for future research would be to incorporate the consensus contact information into established tertiary structure prediction methods such as Rosetta (Simons et al., 1997) in the form of soft constraints to guide sampling procedures or to pick among energetically similar models.

To demonstrate that the contact information can indeed be exploited to improve 3D predictions, we performed another experiment that is based on a very simple selection procedure which we call *closest-to-consensus*. In the Casp context this procedure can be stated as follows:

1. Submit the target sequence to as many prediction servers as possible
2. Collect the predicted models and calculate the sample mean contact map
3. Score each input model by element-wise multiplication of the contact matrix with the matrix representation of the sample mean. This measures intuitively, how close the model is to the consensus.
4. Pick the model with the highest score and submit it as a prediction.

The performance of this procedure is shown in Table 3.4.

Table 3.4: Performance of the *closest-to-consensus* strategy. Shown are the average and median GDT-TS score over all targets. All other methods have lower scores (not shown here)

Score	Closest-to-consensus	Zhang-server
average GDT-TS	58.66	59.09
median GDT-TS	64.86	64.39

Evaluated on the Casp8 data, this method would rank first in terms of median GDT-TS and second in terms of average GDT-TS (see Table) following the best performing method in Casp8 (Zhang server).

This reinforces the presumption that there is potential for improvement also for 3D prediction methods. A state-of-the-art 3D prediction method which would take into account the contact constraints from the consensus, would potentially be superior to other servers also in terms of GDT-TS score.

3.3.3 The CMView Software

The methods presented in this chapter have been implemented as part of a graphical Java application called CMView. The implemented features include visualization of the sample mean, consensus contact prediction as used for the Casp assessment and contact based homology modelling including 3D reconstruction using distance geometry. The software is freely available for Linux, Windows and MacOS and other platforms supporting Java . The source code is available under the GNU General Public License. A comprehensive user's manual describes the features and usage of the program. The manual and the program itself can be downloaded from `http://www.bioinformatics.org/cmview`. A screenshot of a typical working session is shown in Figure 3.5.

3.4 Conclusion

We have presented a new consensus method for the prediction of residue-residue contacts. Applied to benchmark data from the Casp8 experiment the method performs better at predicting contacts than any current structure prediction method. Preliminary results from the blind-test Casp9 assessment confirm that it is the best current contact-prediction method available.

The closest-to-consensus experiment demonstrates that a very simple procedure, which uses the consensus contact information for model picking, is already on par with the best individual 3D prediction methods. By using the contacts in more evolved ways, such as for reducing sampling space or as constraints for template picking, it is likely that improved structure prediction methods can be developed which outperform current methods.

With the CMView software, we provide a user-friendly interface to the methods presented here which can be used for protein modeling or programmatically through the enclosed Java library which is available under the GPL open source license.

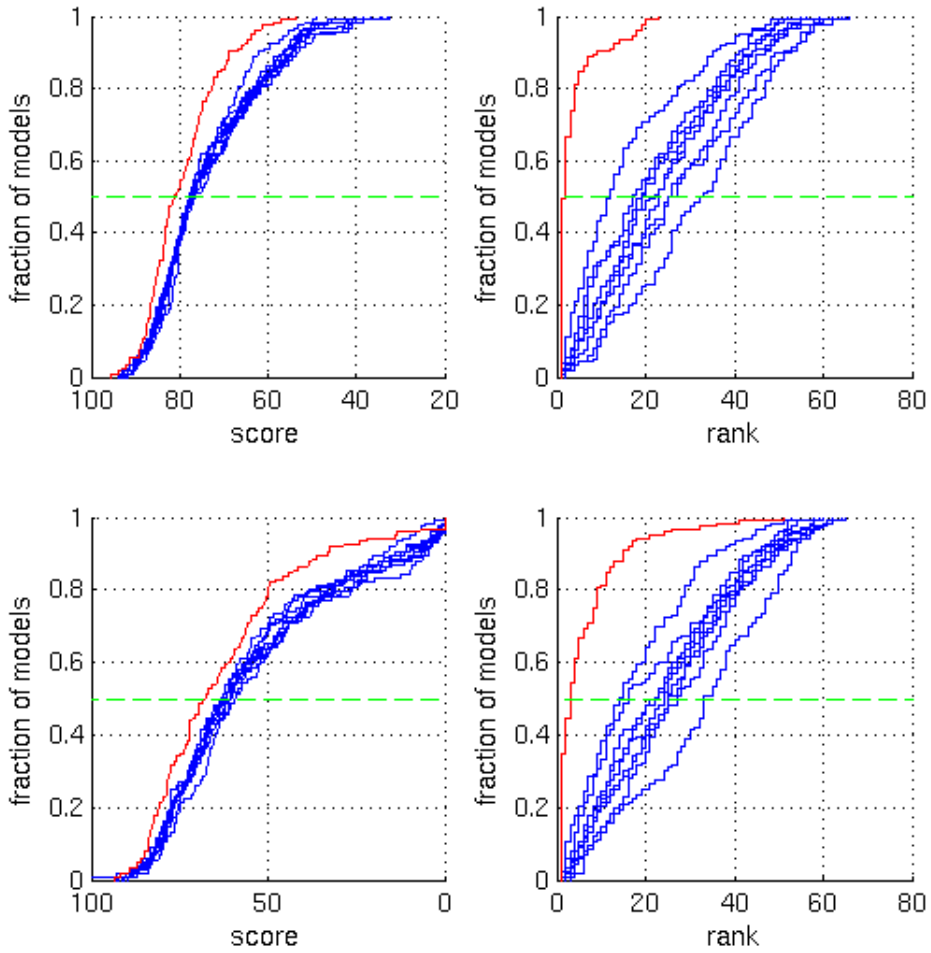

Figure 3.4: Comparison of prediction methods. Our SMEG-CCP method is shown in red. The ten best servers by median CM-score are shown in blue. The curve shows the fraction of models for which the CM-score (a+c) or the rank (b+d) is better than the cutoff on the x-axis. The dashed green line shows the median. The scores in the upper plots (a+b) are evaluated for all contacts, the scores in the lower plots (c+d) for long range contacts only.

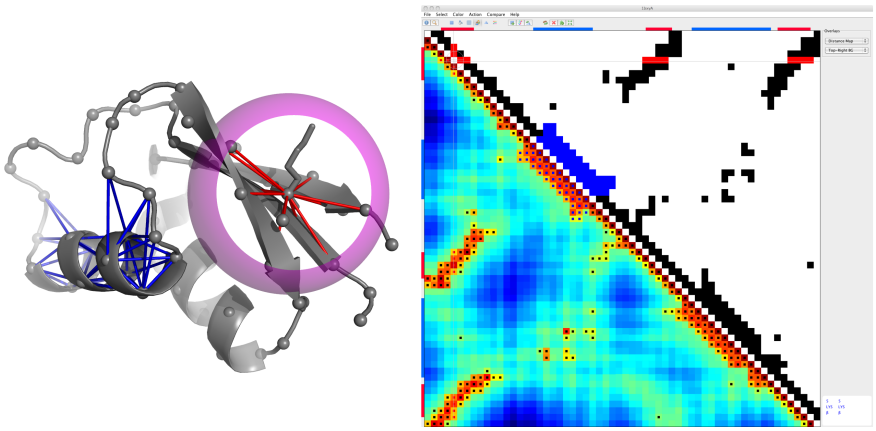

Figure 3.5: Screenshot of the CMView application. The figure shows a typical working session. The application provides two windows, a 3D view of the structure (left) implemented via an interface to PyMol, and the contact map window (right). The two windows are syncronized such that selections in the contact map are automatically shown in the structure. The session shows Ribosomal Protein L30 from Thermus Thermophilus (PDB code 1bxy). The methods for consensus contact prediction and homology modeling described in this chapter are implemented in CMView.

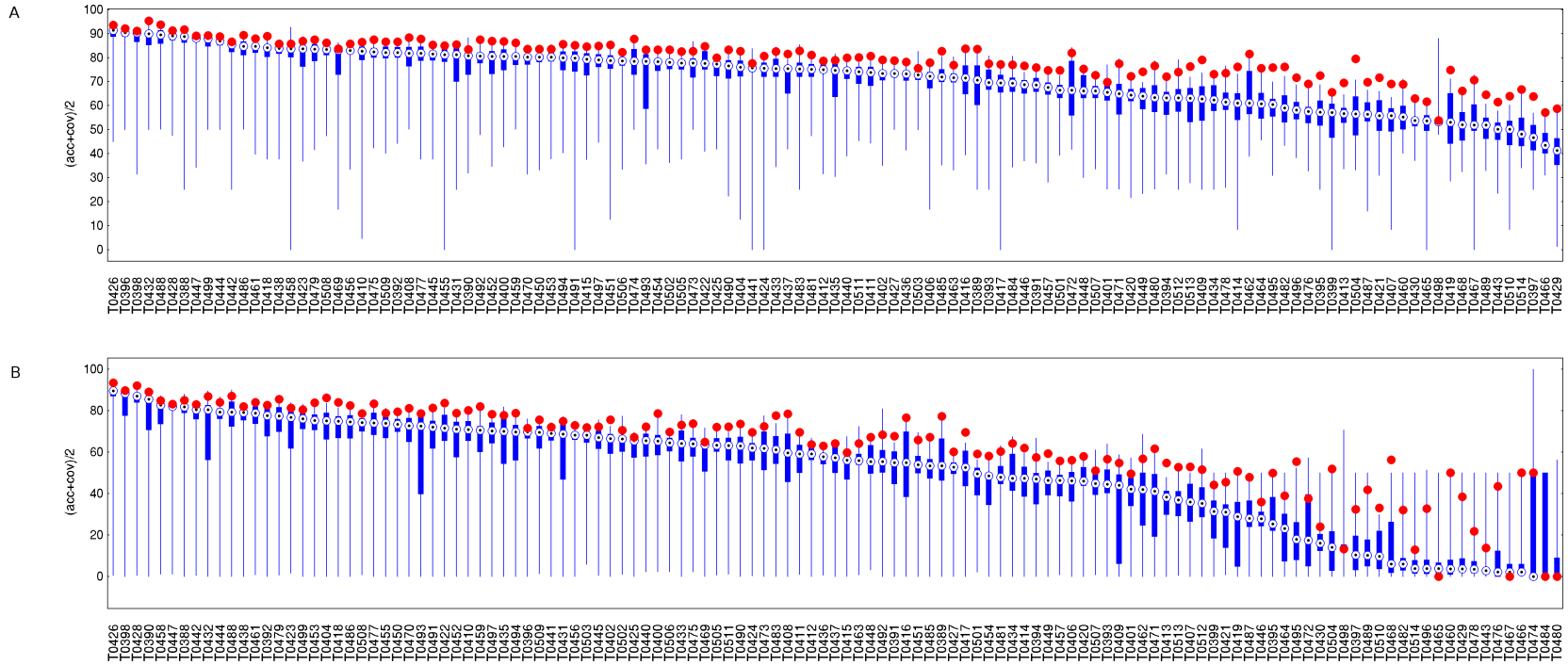

Figure 3.6: SMEG-CCP performance for all Casp8 targets for (A) all contacts and (B) only long range contacts. SMEG-CCP model quality is shown in red. Boxplots show the distribution of the input models with median score shown as white circle and population between 25% and 75% quantile within box. Targets are ordered by median score (≈difficulty).

Chapter 4

Structural consequences of cancer-associated mutations

4.1 Introduction

In recent years, we have witnessed a revolution in genome sequencing technologies. The complete genomic sequence of a human being can now be determined in a matter of days and the costs and processing times are continuously dropping. This progress in technology makes it possible to study the genetic background of individual patients and opens the door for personalized medicine. One of the promising avenues is the analysis of the genomic aberrations that lead to cancer. To this end, several cancer genomics projects have been launched (Wood et al., 2007; Jones et al., 2008; Greenman et al., 2007; Sjoblom et al., 2006; Pleasance et al., 2010a,b; Cancer Genome Atlas Research Network, 2008). By comparing the genomic sequence of cancer cells to healthy tissue of the same patient, somatic aberrations such as point mutations, small insertions or deletions, copy number variations and genomic rearrangements can be identified. These variations promise to provide insights into the functional mechanisms that lead to the uncontrolled cell growth that is characteristic for cancer. However, the step to go from mutation data to functional knowledge remains a challenge.

In this chapter, we will show how we can make use of structural information for the functional interpretation of cancer mutations. We analyzed the effects of ≈2000 cancer-associated missense mutations (i.e. non-synonymous mutations in protein coding regions) on their respective protein structures, and compared the results to the effects of natural variants (SNPs) and randomized mutations. We assessed the effects on the structures with respect to four structural properties: solvent accessibility, protein stability, proximity to functional sites and spatial clustering. To our knowledge, a structural analysis of a similarly large number of cancer mutations has not previously been carried out.

Related work

Some previous studies have analyzed properties of cancer mutations based on sequence features. Such properties include sequence conservation of mutated positions (Hurst et al., 2009), ancestral alleles and substitution propensities (Talavera et al., 2010), and analysis of domain types targeted by mutations (Chittenden et al., 2008). In our analysis, we focus on mechanisms of cancer mutations that have a consequence at the structural level. This makes our method complementary to sequence-based approaches. Another significant body of work has been published on consequences of mutations in a structural context (Ng and Henikoff, 2006, 2003; Ramensky et al., 2002; Wang and Moult, 2001; Karchin, 2009). These studies differ in that either they focus on estimating the severity of individual mutations without looking at specific functions or they use a much broader definition of disease mutations. We will show that by focussing on cancer data and by breaking the set of mutations into more specific subclasses, functionally relevant information is revealed which would be missed otherwise. In particular, we find distinct mutational patterns in oncogenes and tumor suppressors reflecting mechanisms of functional activation and inactivation at the structural level. We statistically validate the observations and show how these differences can be used to predict functional properties of previously uncharacterized genes.

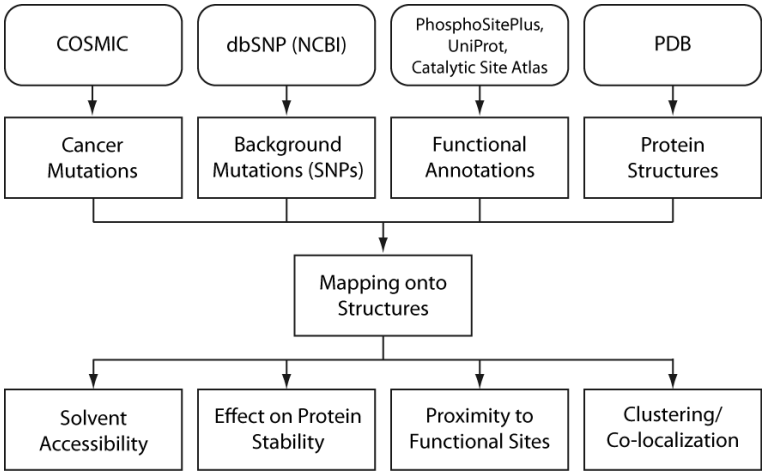

Figure 4.1: Workflow of the analysis

4.2 Methods

An overview of the analysis is shown in Figure 4.1. The cancer mutations as well as functional annotations from public databases are mapped onto the structures, which are taken from the Protein Data Bank (PDB). The effect of each mutation is evaluated with respect to the four structural properties solvent accessibility, protein stability, proximity to functional sites and spatial clustering. A statistical analysis compares the average properties of cancer mutations to those of natural variants from dbSNP and of randomized mutations. In the following, the individual steps are explained in more detail.

4.2.1 Datasets

Cancer mutation dataset (Mut)

Somatic mutations for eight cancer types (breast, prostate, stomach, colon, pancreas, thyroid, kidney, lung) were taken from the COSMIC database (Catalogue of Somatic Mutations in Cancer) (Forbes et al., 2008). For these cancer types, all genes were extracted from COSMIC (v49) for which crystal structures (each of length ≥ 30 amino acids and together covering at least 25% of the gene) were available and which were part of the "Cancer Gene Census" category of COSMIC. For genes in this category a comprehensive literature screening has been conducted. A cutoff of 6 distinct missense mutations for each gene in the structurally resolved regions was chosen based on the observation that genes with very few mutations show high statistical fluctuations. As we exclusively consider missense mutations, we refer to them as "mutations" hereafter. The genes and the corresponding mutations were subsequently separated into the two datasets Onc and Sup representing the subset of mutations in oncogenes and tumor suppressors, respectively (see Table 4.1). A graphical overview of the mutations along the sequence as well as the coverage of the crystal structures is provided in Figures 4.10 and 4.11. The set of genes results from the described automatic selection procedure without any manual intervention.

Single nucleotide polymorphism dataset (Snp)

As a control set, we extracted single nucleotide polymorphism (SNP) data for the 24 genes from version 131 of the common variation database dbSNP (Wheeler et al., 2008). Minor allele frequency data was only available for a small subset of dbSNP entries. Therefore, we excluded those SNPs that are annotated by dbSNP as disease-associated instead.

Random control dataset (Rnd)

As an additional control and as the null-model for the statistical analysis we generated sets of randomized mutations in the 24 genes. These were obtained by randomly permuting the set of mutations in each gene 100 000 times.

4.2.2 Structural Features

Known crystal structures were taken from the Protein Data Bank (Berman et al., 2003). The ones with the largest sequence coverage and with the best crystallographic resolution were chosen.

Structure models of DCLK3 and ERBB2 (see section 4.3.6) were built using an in-house pipeline based on established homology modeling principles. Templates were identified by a PSI-Blast search with 5 iterations (Altschul et al., 1997). Models were built using distance geometry and subsequent simulated annealing refinement.

Table 4.1: Overview of genes used in the analysis

Gene Name	Length (AA)	Mut	SNP	PDB Codes (sequence range)
Oncogenes				
AKT1	478	6	5	1UNQA (1-123), 3CQWA (144-480)
BRAF	766	46	3	3D4QA (433-726), 3NY5A (153-237)
EGFR	1210	224	9	3D4QA (433-726), 3NY5A (153-237)
GNAS	394	12	9	1AZSC (1-394)
HRAS	189	19	0	4Q21A (1-189)
KIT	976	9	9	2EC8A (1-519), 3G0EA (544-935)
KRAS	188	85	1	3GFTA (1-164)
MET	1408	24	30	2UZXB (25-740), 3DKCA (1049-1360)
NRAS	189	9	1	3CONA (1-172)
PIK3CA	1068	148	17	2RD0A (1-1068)
PTPN11	593	7	6	2SHPA (3-529)
RET	1114	24	3	2IVSA (705-1013), 2X2UA (29-270)
Tumor Suppressor Genes				
CDH1	882	17	3	2O72A (155-367)
CDKN2A	156	76	10	1BI7B (1-156)
FBXW7	707	34	4	2OVRB (263-707)
MLH1	756	8	3	3NA3A (1-347)
MSH2	934	12	17	2O8BA (1-934)
PTEN	403	93	2	1D5RA (8-353)
RB1	928	7	9	2R7GA (380-787), 2QDJA (52-355), 2AZEC (829-874)
SMAD4	552	51	3	1DD1A (285-552)
STK11	433	30	1	2WTKC (43-347)
TP53	393	826	17	2VUKA (94-312), 1AIEA (326-356)
VHL	213	216	16	1LM8V (54-213)
WT1	449	9	3	2PRTA (318-438)

Abbreviations: AA, amino acid, MUT, number of mutations, SNP, number of SNPs, PDB, Protein Data Bank

Structural feature - solvent accessibility

Solvent accessibilities were computed using the NACCESS software (Hubbard and Thornton, 1993). NACCESS calculates the relative solvent accessibility (RSA) using a water probe. It implements the probe sphere algorithm illustrated in Figure 4.2. Residues were considered to be solvent accessible or "surface residues" if the RSA was greater than 15% (see Figure 4.3) The odds ratio is calculated as observed over expected fraction of surface mutations in a gene.

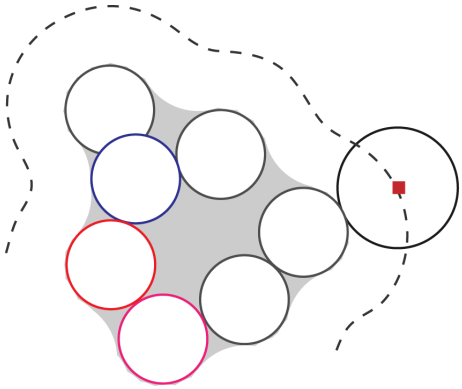

Figure 4.2: Schematic view of the probe sphere algorithm (Shrake and Rupley, 1973). For simplicity, the procedure is illustrated in two dimensions. A solvent probe is rolled around the protein. The trace of the probe center is shown as a dashed line. In three dimensions, the trace of the sphere center describes a surface. The size of this surface is called the *solvent accessible surface area* (SASA). Every point on the surface is assigned to the residue nearest to the probe center to define the residue specific surface area. This per-residue SASA is divided by a SASA value for the respective residue in isolation to calculate the per-residue *relative solvent accessibility* (RSA).

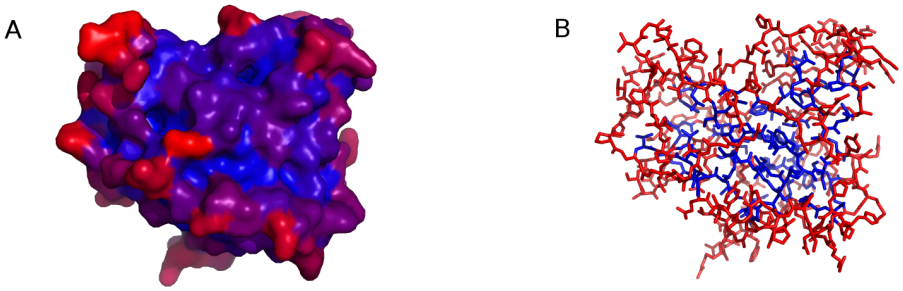

Figure 4.3: Illustration of solvent accessibility. A: The picture shows the protein surface colored by the relative solvent accessibility (RSA) as calculated by NACCESS. The scale ranges from red for completely exposed (100% RSA) to blue for completely buried (0% RSA). B: By applying a cutoff (in this case at 15% RSA), the residues are partitioned into surface residues (red) and core residues (blue). The structure shown is the core domain of human p53 (PDB code 2vukA).

Structural feature - protein stability

To estimate the effect of a mutation on protein stability we used version 3.0 beta of the FoldX software (Guerois et al., 2002). Calculations were performed using the BuildModel command with default parameters. Mutations are considered destabilizing if the difference in free energy between wild type and mutants ($\Delta\Delta G$) exceeds 5 kcal/mol. This value is a typical lower bound for the stability of globular proteins (Chittenden et al., 2008). Otherwise, the mutation is considered neutral. The odds ratio is calculated as observed over expected fraction of destabilizing mutations in a gene.

$$\begin{aligned}\Delta G = {} & W_{vdw} \cdot \Delta G_{vdw} + W_{solvH} \cdot \Delta G_{solvH} + W_{solvP} \cdot \Delta G_{solvP} + \Delta G_{wb} \\ & + \Delta G_{hbond} + \Delta G_{el} + \Delta G_{Kon} + W_{mc} \cdot T \cdot \Delta S_{mc} + W_{sc} \cdot T \cdot \Delta S_{sc}\end{aligned}$$

$\Delta G_{solvP/H}$	difference in solvation energy for polar/apolar groups
ΔG_{vdw}	sum of van der Waals contributions
ΔG_{wb}	extra stabilizing free energy of water bridges
ΔG_{hbond}	energy difference of intra-molecular H-bond formation
ΔG_{el}	electrostatic contribution of charged groups
ΔS_{mc}	entropy cost of fixing the backbone in the folded state
ΔS_{sc}	entropic cost of fixing a side chain in a particular conformation
ΔG_{kon}	effect of electrostatic interactions on the association constant k_{on}

Figure 4.4: Energy function for calculating protein stability used by FoldX. The stability ΔG is the difference in free energy between folded and unfolded state. ΔG is being evaluated before and after mutation. The difference $\Delta\Delta G$ is then the stability change upon mutation which is used to characterize individual mutations. Further details about the energy function can be found in (Guerois et al., 2002).

Figure 4.5 shows a histogram of the destabilizing effect of the population of all possible mutations for our gene data set calculated with FoldX. As expected, most mutations are neutral or even stabilizing. This can be seen as a sanity check for the stability calculations.

Structural feature - proximity to functional sites

A mutation is considered proximal to a functional site if it occurs at or in contact with a functional residue where contact is defined as the C-beta atoms of the respective residues being no more than 8Å apart (C-alpha for glycine). Functional site annotations were derived from public databases (UniProt release 2010-10 (The UniProt Consortium, 2010), Catalytic Site Atlas version 02.02.12 (Porter et al., 2004), PhosphoSitePlus as of 2010-10-15 (Hornbeck et al., 2004)). We extracted the following categories of functional site annotations: Enzyme active sites, ATP/GTP binding

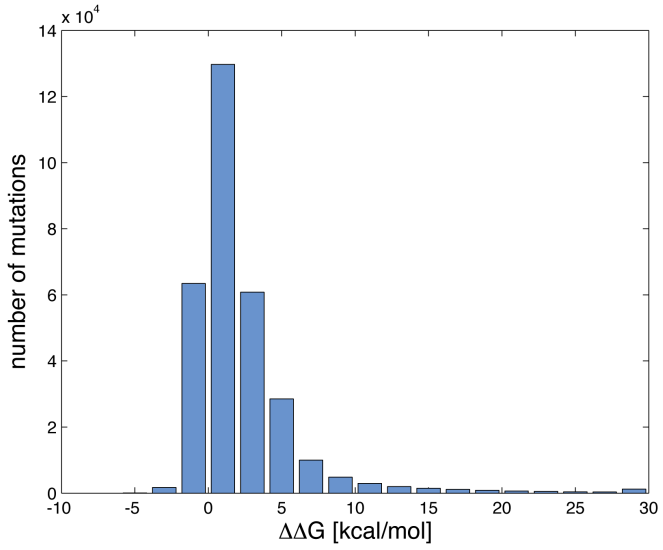

Figure 4.5: Histogram of stability changes for all possible mutations. The plot shows few mutations that are stabilizing ($\Delta\Delta G < 0$), the majority of mutations which are mildly destabilizing and a long tail of severely destabilizing mutations ($\Delta\Delta G > 5$).

sites, phosphorylation sites, ubiquitination, and other post-translational modifications (acetylation, methylation, and glycosylation). The odds ratio is calculated as observed over expected fraction of mutations proximal to a functional site.

Structural feature - spatial clustering

To measure whether a set of mutations is spatially clustered in the structure, we divide the protein into structurally defined domains and calculate a spatial clustering value C as follows:

$$C = \frac{1}{N} \sum_{i,j} \frac{1}{d_{i,j}} \tag{4.1}$$

where $d_{i,j}$ is the Euclidean distance in the structure between the side chain centroids of residues i and j, and N is the number of such residue pairs. We used the C-alpha position for glycines and residues with unavailable side chain coordinates. The domains are structurally defined using the DomainParser method (Xu et al., 2000). Only distances within domains are evaluated. The subdivision into domains is crucial to avoid bias due to the size and domain architecture of the protein. The odds ratio is calculated as observed over expected clustering value of the mutations in a gene.

4.2.3 Statistical Analysis

Odds ratios

The structural features described below are evaluated for each gene in terms of the odds ratio of observed over expected behavior. Expected values are calculated by generating

a large population of randomized sets of mutations and evaluating the property (e.g. fraction of solvent accessible residues in the structure) averaged over the population.

P-values

The statistical significance of the observations was assessed by calculating the p-value under the null-model assumption of a uniform distribution of the mutations. In the cases with a binary outcome for each position (surface/core, neutral/destabilizing, proximal to functional site or not), the null-model distribution is binomial and the p-value can be calculated analytically. Otherwise, it has to be obtained by simulation (spatial clustering). Let f be such an empirical null-model distribution with mean m. Then, the p-value of an observation O is approximated as the fraction of individuals v in the population with $f(v) \geq f(O)$ if $O \geq m$ or $f(v) \leq f(O)$ if $O < m$.

Jackknife test

To assess the robustness of the data against outliers, we applied a jackknife test. This test is a bootstrapping procedure where the results are being recalculated multiple times, each time leaving out one gene from the original dataset. Taking the maximum and the minimum over this procedure for all genes yields an interval around the value of the original dataset. These intervals are shown as error bars in the Figure 4.6.

Linear classifiers

Linear classifiers were automatically calculated using Fisher's linear discriminant method, which provides a good compromise between finding the optimal solution in the linearly separable case and being robust to outliers (Fisher, 1936). To test the robustness of the classification we applied a leave-one-out cross validation procedure. In each step, one gene is temporarily removed from the training set. The classifier is recalculated on the subset and we test whether it is able to correctly predict the class membership of the excluded gene.

4.3 Results

In this study we analyzed the structural impact of a large number of cancer mutations in oncogenes and tumor suppressors. We evaluated the impact with respect to four structural features. We focused on eight selected tumor entities that are among the most frequent and lethal types. The Mut dataset extracted from the COSMIC database (Forbes et al., 2008) comprises 1992 mutations in 24 cancer genes. This set contains many classical cancer genes that are involved in major signaling pathways (i.e. TGFb, EGFR, MAPK, PI3K/AKT signaling). The genes with their corresponding mutations were subdivided into the classes of tumor suppressor (Sup) and oncogenes (Onc) as shown in Table 4.1, representing two common mechanisms through which

tumorigenesis is initiated: via gain-of-function of oncogenes and loss-of-function of tumor suppressors (Vogelstein and Kinzler, 1993). As a control, we used a set of 204 non-disease-related SNPs (the Snp dataset) extracted from NCBI's database dbSNP (Sherry et al., 2001). In the following, we present the results for the four structural properties. The reported values in Figure 4.6 are the average odds ratios over the genes in the respective set (Snp, Mut, Onc, Sup).

4.3.1 Solvent accessibility

As the first property, we investigated whether mutations occur at the surface or in the core of the protein. Figure 4.6A shows that there is little difference between the SNPs (Snp, 0.938) and cancer mutations (Mut, 0.987). However, a separate analysis of oncogenes and tumor suppressors reveals that mutations in oncogenes occur significantly more often at the surface (1.122, p-value 2e-5), while mutations in tumor suppressors are overrepresented in the core (0.852, p-value 4.4e-16).

4.3.2 Protein stability

We calculated the impact that the mutations of the different datasets have on protein stability. The calculations were performed with the FoldX software (Guerois et al., 2002). A recent assessment has shown that this method is currently among the best methods for calculating stability changes upon mutation (Potapov et al., 2009). The results of this analysis (Figure 4.6B) show a distinct difference between oncogenes and tumor suppressors. Tumor suppressors display a significant overrepresentation of mutations that destabilize the protein (1.903, p-value 2.9e-11) with an almost four-fold increase compared to oncogenes with significantly fewer destabilizing mutations (0.513, p-value 5.8e-7).

4.3.3 Proximity to functional sites

Next we assessed whether the mutations in our dataset occur proximal to known functional sites and thus are likely to directly influence protein function. For this we extracted 258 annotated functional sites from public databases. The results are shown in Figure 4.6C. Cancer mutations in oncogenes (Onc) have a tendency to specifically target functional sites (1.663, p-value 1e-5), while in tumor suppressors (Sup) mutations proximal to functional sites are significantly underrepresented (0.893, p-value 4.6e-2). Functional site mutations are also significantly underrepresented in the Snp data set (0.770, p-value 4.4e-8). Further, we investigated whether particular types of functional sites are more often mutated than expected. Figure 4.7 shows the observed distribution of functional site mutations in oncogenes and tumor suppressors compared to the distribution expected for randomized mutations. For oncogenes, ATP and GTP binding sites are significantly overrepresented among the mutated functional sites (31% compared to 16%, p-value 4.95e-11 (ATP) and 22% compared to 13%, p-

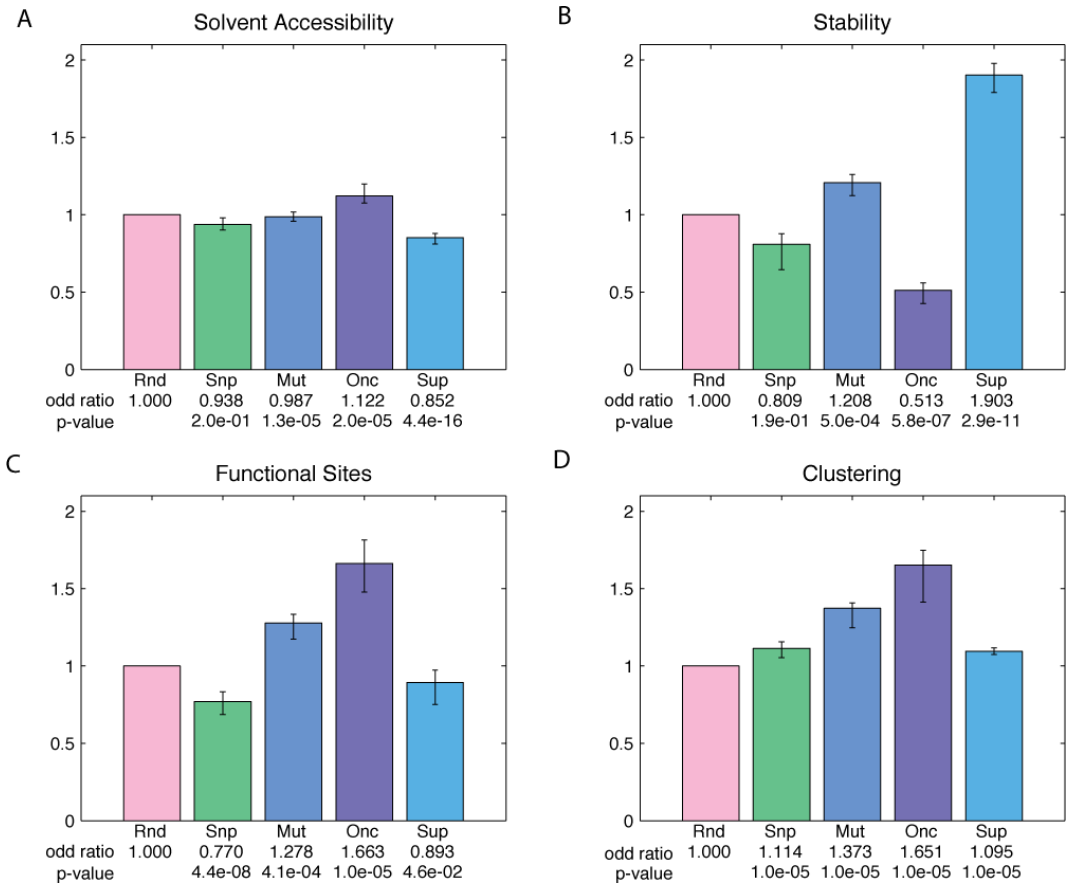

Figure 4.6: Structural impact of mutations. The columns show the structural properties of random mutations (Rnd), natural variations (Snp) and cancer mutations (Mut). Cancer mutations are further analyzed separately as mutations of oncogenes (Onc) and mutations of tumor suppressor genes (Sup). The error bars indicate the variability of the data under the jackknife test. A, observed over expected fraction of mutations occurring at the protein surface. Onc show significantly more and Sup significantly less solvent accessible mutations. B, observed over expected fraction of destabilizing mutations. Onc mutations are less often destabilizing, while Sup mutations disrupt stability far more often than the controls. C, observed over expected functional site mutations. Functional sites are more frequently mutated in Onc than in Sup. D, observed over expected spatial clustering of mutations. Mutations particularly in Onc are significantly more clustered than expected by chance.

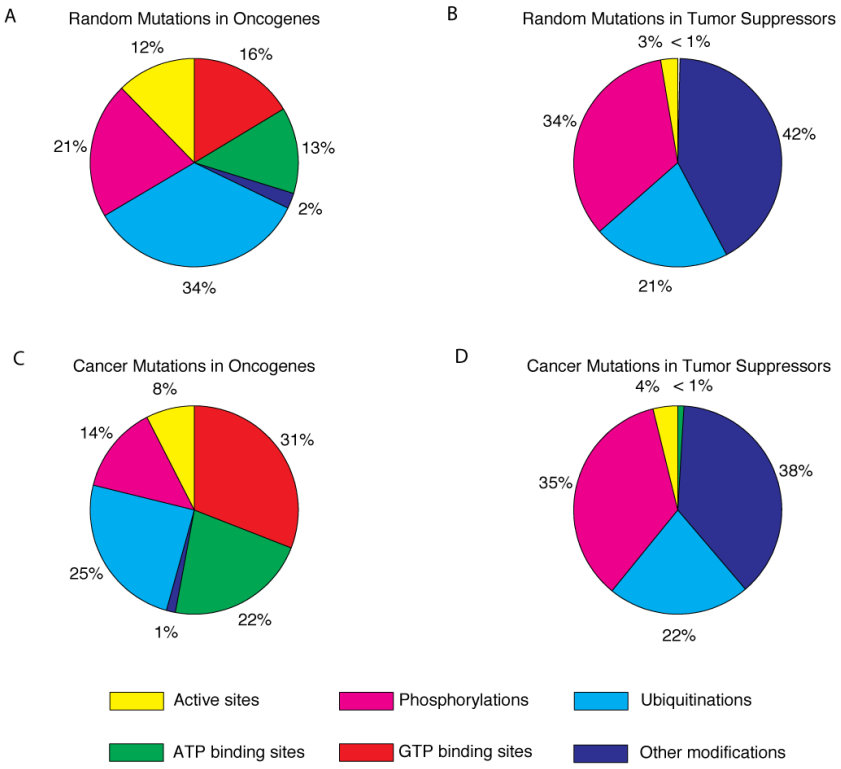

Figure 4.7: Distribution of functional site mutations. Distribution of mutations affecting functional sites in oncogenes (Onc) and tumor suppressors (Sup) compared to distribution of random mutations. A and B, distribution obtained by random sampling of positions in Onc and Sup, respectively. C, distribution of functional site mutations in Onc. ATP and GTP binding sites in Onc are significantly more often mutated than expected by chance. D, distribution of functional site mutations in Sup. Observed distribution does not differ significantly from expected random distribution.

value 4.86e-07 (GTP)). The results for tumor suppressors show no apparent differences between observed and random distribution.

4.3.4 Spatial clustering

Next, we wanted to test whether cancer mutations have a tendency to co-localize in spatial clusters. Figure 4.6D shows that cancer mutations in oncogenes are highly clustered (1.651), while tumor suppressor mutations behave similar to SNPs (1.095 compared to 1.114). Both are significantly more clustered than random (p-value $< 1e-5$). The small error bars for Sup indicate that all tumor suppressors have similar clustering behavior. In this case, the p-values results from the fact that a spatial clustering as high as the one for either of the sets Snp, Mut, Onc or Sup was never observed in the random reference population of size 100 000. Hence, the p-value is at most 1e-05.

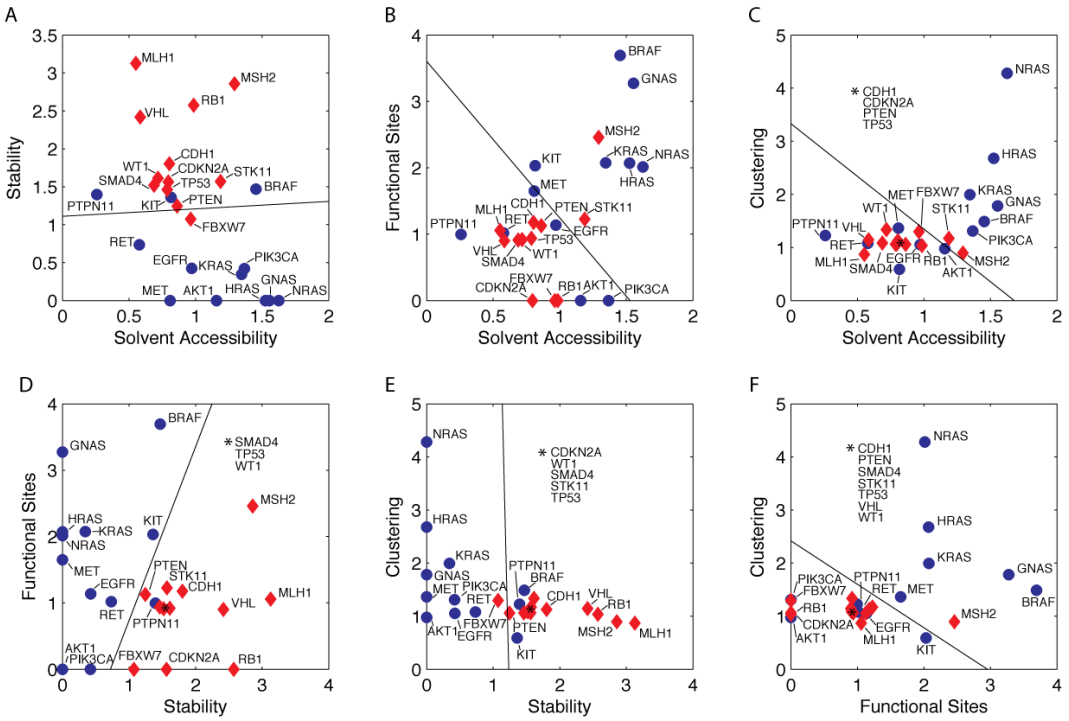

Figure 4.8: Linear classification of cancer genes. The different pairs of structural features are shown as scatter plots in A-F. Oncogenes are depicted as blue dots, tumor suppressors as red diamonds. The separating linear functions have been calculated using Fisher's linear discriminant method. The classifiers in A, D and E show the best training performance.

4.3.5 Classification of cancer genes based on structural features

Given the distinct average behavior of the two cancer gene classes, we investigated to what extend this behavior is reflected at the individual gene level and to what extend it can be used for predictive purposes. To examine the discriminatory power of the structural features, the features were plotted in pairwise combinations (Figure 4.8). Each data point corresponds to one individual gene with oncogenes and tumor suppressors are shown as blue dots and red diamonds, respectively. The values on the axes are the odds ratios for the feature values. We calculated linear classifiers trained on the two sets using Fisher's discriminant method (Fisher, 1936). Visually, the two classes are well-separated for feature combinations shown in Figure 4.8A, 4.8D and 4.8E. For combinations in Figure 4.8B, 4.8C and 4.8F, the two subpopulations overlap more. Nevertheless, in all six plots there are areas exclusively populated by either class. We have systematically evaluated the discriminatory power of the different feature combinations (see Table 4.2) by leave-one-out cross validation. We find that the combination of the two features functional sites and stability (Figure 4.8D) classifies best with a performance of 95.83%. The plots in Figure 4.8A and 4.8E (stability vs. surface accessibility, clustering vs. stability) display a cross validation performance of 83.33% and 79.17%, respectively. The other feature combinations possess modest classification power.

Table 4.2: Performance of linear classifiers and prediction of functional classes - linear classifiers were calculated for each combination of the four structural features. Shown are the classification performance on the training set and in a leave-one-out cross validation. The classifiers were applied to predict functional classes for five cancer genes.

Feature Combination				Training Performance			Cross Validation			Production					
1 feature				true	false	ratio	true	false	ratio	DCLK3	MMP2	PIK3C3	TGM3	ERBB2	EPHA3
Surf				17	7	70.83%	16	8	66.67%	O	S	S	O	O	S
Stab				20	4	83.33%	20	4	83.33%	O	S	S	O	O	S
Func				18	6	75.00%	18	6	75.00%	S	O	S	S	S	S
Clust				17	7	70.83%	17	7	70.83%	S	O	S	S	S	S
2 features															
Surf	Stab			20	4	83.33%	20	4	83.33%	O	S	S	O	O	S
Surf	Func			16	8	66.67%	16	8	66.67%	S	O	S	O	S	S
Surf	Clust			16	8	66.67%	16	8	66.67%	S	O	S	O	S	S
Stab	Func			23	1	95.83%	23	1	95.83%	O	O	S	O	O	S
Stab	Clust			20	4	83.33%	19	5	79.17%	O	S	S	O	O	S
Func	Clust			18	6	75.00%	17	7	70.83%	S	O	S	S	S	S
3 features															
Stab	Func	Clust		23	1	95.83%	23	1	95.83%	O	O	S	O	O	S
Surf	Func	Clust		18	6	75.00%	15	9	62.50%	S	O	S	S	S	S
Surf	Stab	Clust		20	4	83.33%	19	5	79.17%	O	S	S	O	O	S
Surf	Stab	Func		24	0	100.00%	22	2	91.67%	O	O	S	O	O	S
4 features															
Surf	Stab	Func	Clust	24	0	100.00%	22	2	91.67%	O	O	S	O	O	S
Consensus										O	O	S	O	O	S
Bozic et al.										O	O	-	-	O	O

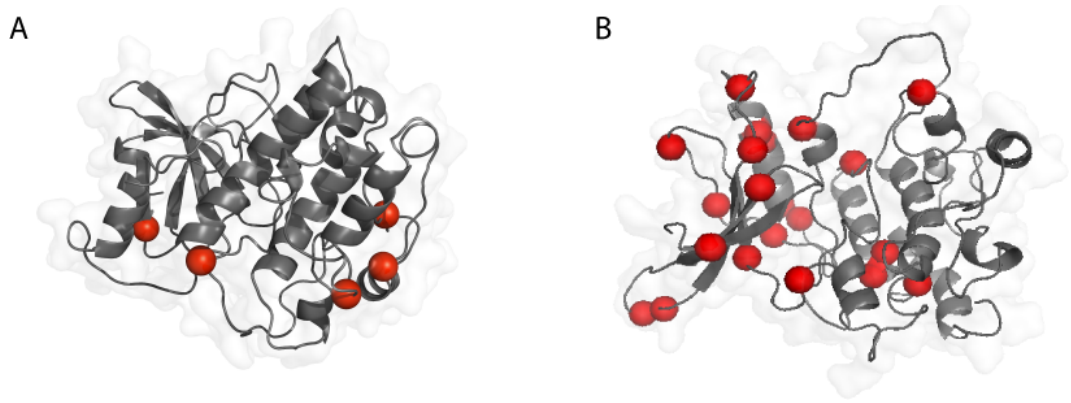

Figure 4.9: Structural models for DCLK3(A) and ERBB2(B)

4.3.6 Prediction of gene function

Given the good performance of the classifiers, we applied the classification to six genes with uncertain annotation (MMP2, PIK3C3, TGM3, EPHA3, DCLK3, ERBB2). These genes were not included in our original dataset either because they were not in the "Cancer Gene Census" category of COSMIC (MMP2, PIK3C3, TGM3, EPHA3) or because there was no crystal structure available (DCLK3, ERBB2). We generated homology models for DCLK3 and ERBB2 (see Figure 4.9). For EPHA3 we found clear evidence in the literature that it acts as a tumor suppressor (Lee et al., 2010). For the other genes, the classification is less clear. We systematically applied the linear classifiers shown in Figure 4.8 to this set. Table 4.2 shows a summary of the results. The consensus of the classifiers identifies DCLK3, MMP2, TGM3 and ERBB2 as oncogenes and PIK3C3 and EPHA3 as tumor suppressors. This matches the prediction result of the best performing classifier (functional sites versus stability).

4.4 Discussion

4.4.1 Comparison with related work

Previous studies of structural effects of mutations have found that disease mutations primarily occur in the protein core (Wang and Moult, 2001; Ramensky et al., 2002). We can confirm this trend only for the set of tumor suppressors. In contrast, core residues in oncogenes are significantly less often mutated than expected by chance. This is in agreement with our results for protein stability. Mutations located in the protein core are often destabilizing and result in loss-of-function. Thus, our data suggests that the loss-of-function of tumor suppressors is often caused by destabilization of the protein. Similar to our findings, Gong and Blundell show that cancer mutations are less often located in solvent inaccessible areas than expected, as opposed to Mendelian disease-related variants (Jeffers et al., 1997). In another recent study, Talavera et al. report that cancer driver mutations are more likely located on the surface of proteins than expected by chance (Talavera et al., 2010). Their observation that the patterns of cancer associated mutations and common polymorphisms are "remarkably similar"

can be explained by our results that the opposing trends of tumor suppressors and oncogenes neutralize each other when looking at cancer mutations in general.

Functional site mutations can either disable enzymatic activity and regulatory mechanisms or increase protein activity, as it has been described for several examples. One example is the well-characterized V600E mutation in BRAF that mimics the phosphorylation of the kinase domain activation segment (Davies et al., 2002). For the Onc set we observed a significant overrepresentation of mutations proximal to functional sites. This suggests that specific mutations of functional sites are often responsible for oncogene activation. The underrepresentation of functional site mutations in the Snp dataset can be explained by the fact that SNPs are assumed to occur in the population without causing severe phenotypes. A mutation of a functional site impairing the native protein function would be unfavorable. Our results show that the most frequently mutated types of functional sites in oncogenes are ATP and GTP binding sites and that the frequency of mutation is significantly higher than expected. This suggests that mutations of ATP and GTP binding sites are specific and common mechanisms of oncogene activation. In fact, examples for such activating mutations near ATP binding sites have been described in the literature (Davies et al., 2002; Shu et al., 1990; Jeffers et al., 1997). This is supported by previous findings showing that the functional region of ATP binding is subject to a greater selection pressure indicative for the presence of candidate driver mutations (Torkamani and Schork, 2008), and that in kinases this site shows a higher proportion of driver mutations compared to the remaining catalytic domain (Greenman et al., 2007). Further, mutations in the GTP binding site of RAS genes have been described to impair GTPase activity. These mutations retain the protein in a GTP-bound state leading to constant activation of the gene (Malumbres and Barbacid, 2003; Pai et al., 1989). We have observed highly significant spatial clustering of mutations in particular in oncogenes. Similar trends have been described in recent publications (Yue et al., 2010; Ye et al., 2010). Even though different, sequence-based definitions of clustering were used, the results, like ours, support the hypothesis that mutations in specific regions in the structure are required for gene activation. Our results further indicate that tumor suppressor deactivation is a locally less constrained process.

4.4.2 The role of driver and passenger mutations

To identify tumor-causing mechanisms from sequencing data, it is important to distinguish between driver and passenger mutations. By definition, driver mutations are actively involved in the process of tumor formation. In contrast, passenger mutations occur by chance and do not confer any growth advantages. Typically, cancer genomics studies will include a step to filter out passenger mutations and several approaches for such filtering have been described (Greenman et al., 2007; Torkamani and Schork, 2008; Carter et al., 2009; Kaminker et al., 2007). We have only included genes that are taken from the "Cancer Gene Census" part of the COSMIC database and we make the assumption that mutations described in the literature are less likely to be passengers. Nevertheless, there is the possibility that the Mut dataset contains passenger mutations. We expect that they behave more similar to the control sets (Rnd and Snp) and shift the results towards the expected random value. Since the observed

differences between Onc and Sup are so significant, we conclude that the signal from driver mutations dominates the noise induced by passengers. Figure 4.8 shows the behavior of individual genes and the linear classifiers that we trained on the dataset. We find that plots with the stability feature on one axis (Figure 4.8A, 4.8D, 4.8E) show good separation. We looked at some outlier genes with unexpected behavior in more detail. For example, the value for functional site mutations in PIK3CA is zero. This is because the databases were missing annotations described in the literature for the ATP binding- and catalytic sites (Huang et al., 2008). So there is some effect of database contents, but the other genes in our dataset seem to be well-annotated. The two recurring outliers, PTPN11 and AKT1 are the genes with the least number of distinct mutations in our dataset. Therefore, we suggest that results for genes with few mutations should be handled with care and that for a robust classification more mutations are advantageous. Plots involving clustering (Figure 4.8C, 4.8E, 4.8F) show that all tumor suppressors have a similar clustering value around one, whereas oncogenes show a wider distribution with very high and some low values. The three members of the RAS family show the highest clustering values due to the dominance of mutations around the common hotspot at position twelve. KIT shows the lowest clustering value because it is only rarely mutated in the eight selected tumor types and the mutations are even more scattered in the structure than random.

4.4.3 Prediction for novel genes

The results of the cross validation showed good performance of the features for predictive classification. Hence, we used the classifiers to predict the functional class of five genes not included in the original dataset. We compared the predictions of our linear classifiers to recent results by Bozic et al. (Bozic et al., 2010). They conducted a classification of all genes contained in COSMIC into oncogenes and tumor suppressors based on non-structural features. For two of the genes (DCLK3 and MMP2), their classification as oncogenes matches ours. For EPHA3, the two annotations disagree. Our classification is in accordance with prior knowledge about the tumor suppressor activity of EPHA3 (Lee et al., 2010). Further investigations may be required to elucidate this apparent disagreement. For two previously uncharacterized genes (PIK3C3 and TGM3), for which Bozic and coworkers do not report annotations, we suggest that they act as tumor suppressor and oncogene, respectively.

4.5 Conclusion

We have shown how structural information can be used for the functional interpretation of cancer mutations, thereby bridging the gap between sequence data and functional knowledge.

The central contribution of this work is that it describes for the first time in a quantitative way, the opposing structural effects of cancer-associated missense mutations in oncogenes and tumor suppressors. Our findings confirm and statistically validate the

hypotheses for the gain-of-function and loss-of-function mechanisms of oncogenes and tumor suppressors, at the structural level.

Moreover, we have presented a method that can be used to predict whether a newly identified gene likely acts as an oncogene or a tumor suppressor. The method uses structural features that, in lack of experimental structures, can be derived from predicted models. Because we have focused on properties of cancer mutations that act at the structural level, our results are complementary to those from sequence-based methods. This prediction of functional properties based solely on information which can be obtained from tumor sequencing will be a valuable tool in cancer research as more and more cancer genomic data becomes available.

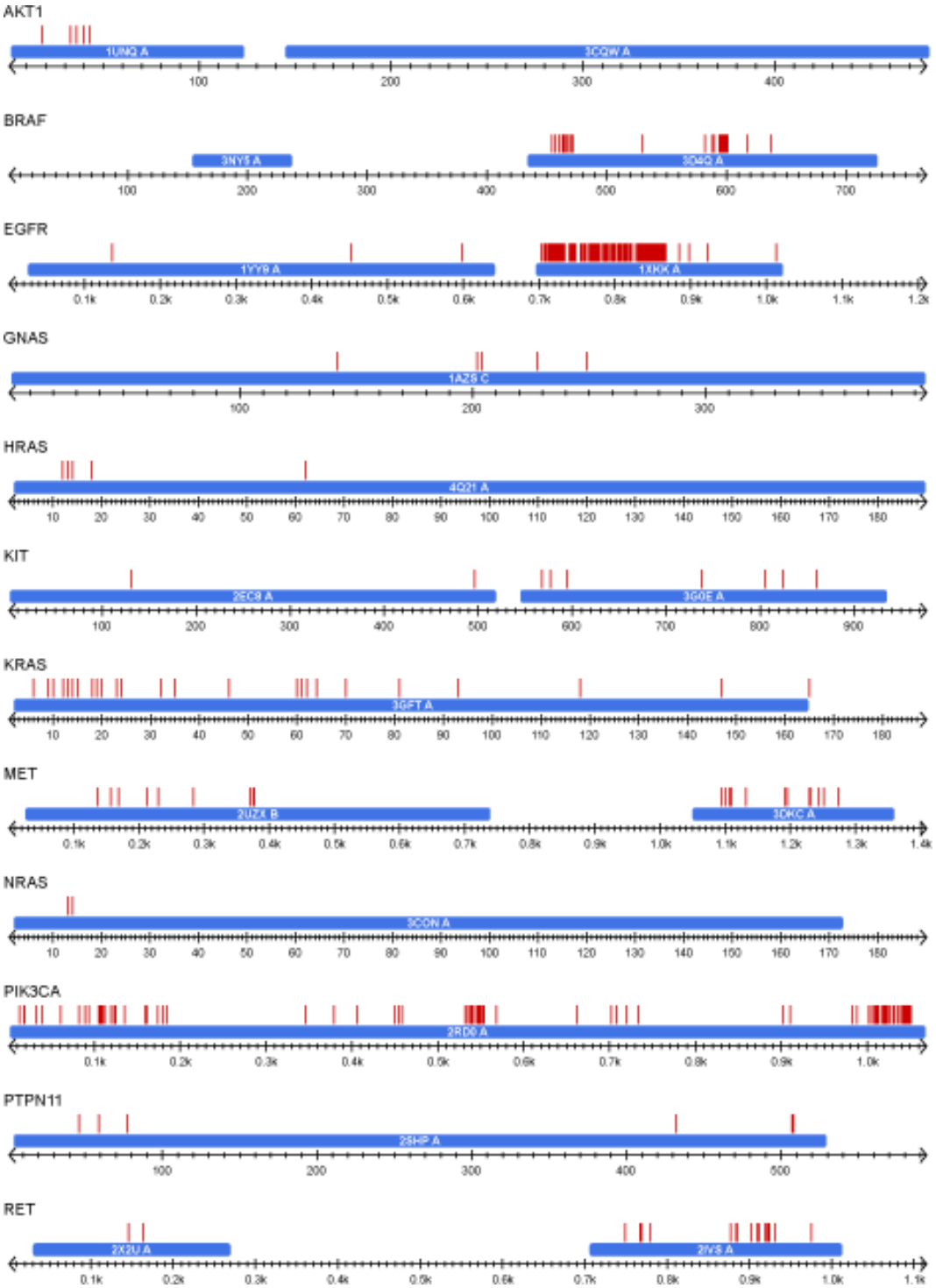

Figure 4.10: Overview of structural regions and locations of cancer mutations for the set of oncogenes uses in the analysis. The structural regions are shown as blue bars labeled with the PDB code. Mutations are shown in red.

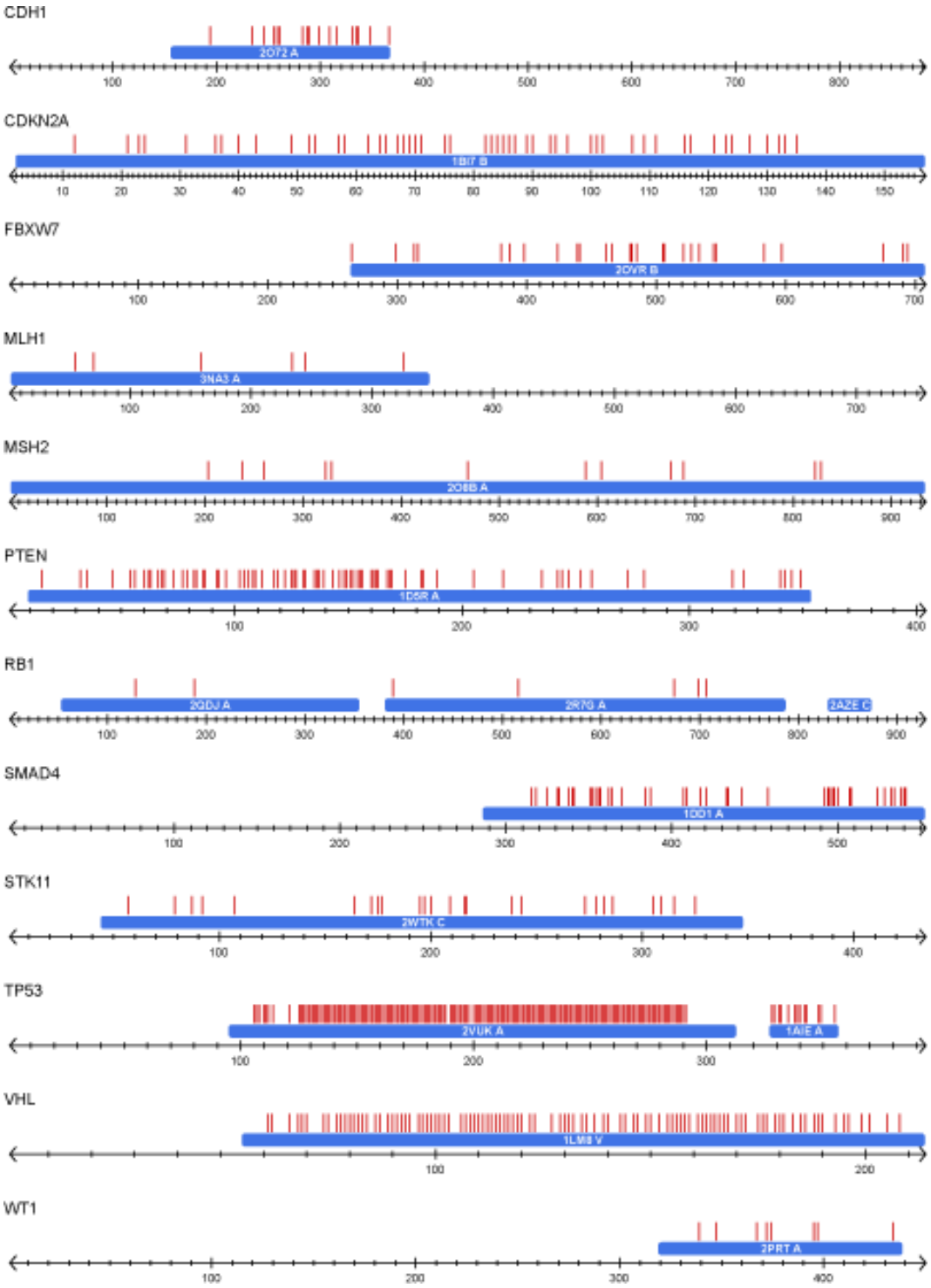

Figure 4.11: Overview of structural regions and locations of cancer mutations for the set of tumor suppressors uses in the analysis. The structural regions are shown as blue bars labeled with the PDB code. Mutations are shown in red.

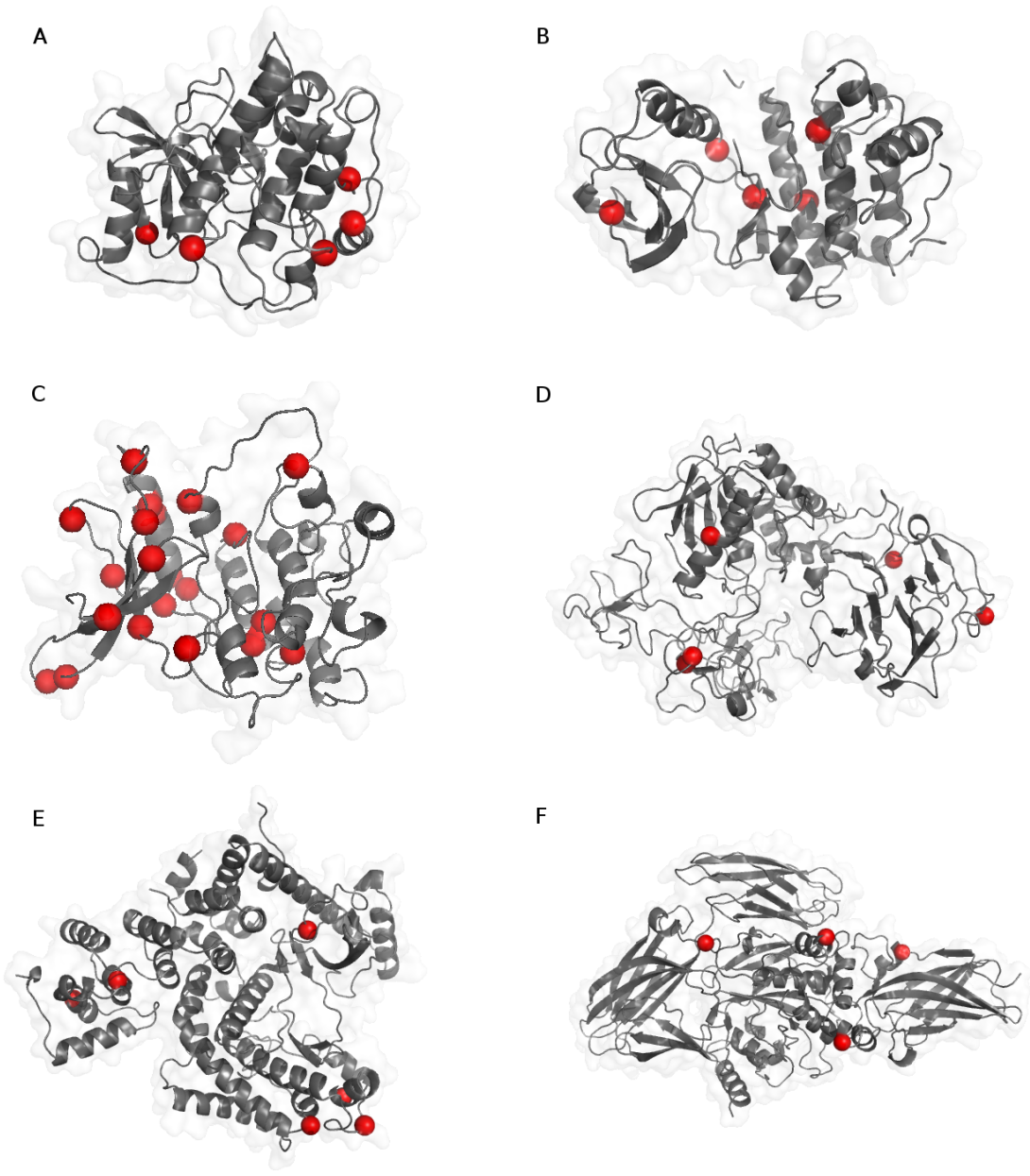

Figure 4.12: Overview of genes for which cancer gene class was predicted. Mutations are shown as red spheres in the structure. A: DCLK3 (Model), B: EPHA3 (PDB 2qolA), C: ERBB2(Model), D: MMP2(1ck7A), E: PIK3C3(3ls8A), F: TGM3 (1vjjA)

Chapter 5

Summary and conclusion

The problem of predicting the tertiary structure of a protein from its sequence has been studied for more than 50 years. Despite significant improvements, automatic methods for structure determination can still not keep up with the pace of generation of other genome-wide data.

Here we have presented novel graph-based methods for protein structure comparison and protein structure prediction. We have further shown how the knowledge of structure can be used to gain insights into mechanisms of human disease.

In Chapter 2, we have shown how to generalize the alignment of contact maps to the multiple case. This allows to define multiple structure alignment (MStA) in a mathematically rigorous way based on maximizing the number of shared contacts. We have shown that in this form, MStA is equivalent to calculating the sample mean of a set of graphs. The main practical implication of this theoretical result is that algorithms developed for finding the sample mean can now be applied to MStA. We have adapted one such method to the multiple structure alignment problem and compared it to other existing methods. We have seen that our method performs well when compared to classical multiple structure alignment methods. Compared to other graph based alignments, which provide similar advantages, our method shows an excellent tradeoff between solution quality and speed.

In Chapter 3, we have presented a new consensus-based method for the prediction of residue-residue contacts. Applied to benchmark data from the Casp8 experiment, the method performs better at predicting contacts than any current structure prediction method. Preliminary results from the blind-test Casp9 assessment indicate that it also compares favorably with dedicated contact-prediction methods. We have shown that a very simple procedure which uses the consensus contact information for model picking is already on par with the best individual 3D predition methods. This shows that even though many state-of-the-art methods already make use of consensus information for template picking and model selection, consensus information at the contact level can be further exploited to improve current prediction methods.

With the CMView software we provide a user-friendly interface to the methods presented here which can be used for protein modeling or programmatically through the enclosed Java library. Both of these tools are available as Open Source Software.

In Chapter 4, we have shown how knowledge of the structure can be used to gain insight into disease mechanisms. Our analysis of ≈2000 cancer-associated mutations describes for the first time in a quantitative way, the distinct mutational effects on oncogenes and tumor suppressors at the structural level. With our findings we can confirm and statistically validate the hypotheses for the gain-of-function and loss-of-function mechanisms of oncogenes and tumor suppressors, respectively. Moreover, we have shown that the different mutational patterns can be used to predict whether a newly identified gene likely acts as an oncogene or a tumor suppressor. This method can be a valuable tool in cancer genome analysis.

Conclusion

In this work, we have addressed three fundamental problems in computational structure biology: Multiple structure alignment, prediction of the tertiary structure from sequence, and the analysis of the structural consequences of mutations. In each case, the main contribution is of a theoretical nature. In Part 1, we discovered the connections between two seemingly unrelated problems, the sample mean of graphs and the alignment of protein structures. The relevance of this result goes beyond the proof-of-principle alignment algorithm we proposed. The main purpose of the algorithm is to demonstrate that methods developed for the sample mean theory can be applied to structure alignment and can compete with previously proposed structure alignment methods. More importantly, the connections to the sample mean and the body of work associated with it, provide a starting point to better understand the theory that is underlying structure alignment. A theory that has been, as Caprara et al describe it "almost nonexistent, as the problems are a blend of continuous-geometric and combinatorial-discrete mathematics" (Caprara et al., 2004).

In Part 2, the main contribution is the demonstration that current structure prediction methods could be improved by making use of consensus information at the level of individual residue-residue contacts. The strategy we propose can certainly be refined but it already predicts contacts better than any of the methods competing in the latest Casp experiments. We can assume that these methods represent the current state-of-the-art of structure prediction methods.

In Part 3 of the work, we provide insights into the effects of cancer-associated mutations on the structures of oncogenes and tumor suppressors. We also show how this gives rise to a method which can predict functional information about a protein in a disease context solely based on the positions of observed mutations. Such mutation data is currently generated in large quantitites by cancer genomics projects.

Taken together, we have shown how to translate sequence information through structure to biological knowledge. The insights and tools presented in this thesis will be helpful for the analysis and interpretation of protein structures and individual mutations in the context of protein structure and function. In an era where the generation of genomic data is not anymore a limiting factor, automatic tools for such interpretation and functional analysis will become ever more important.

Appendix A

Proofs for Chapter 2

Let $\mathcal{X}_n$ be the set of all weighted graphs of order n. We can regard any weighted graph X of order $m < n$ as a graph of order n by appending $p = n - m$ isolated gaps into its vertex set. Thus, we can regard $\mathcal{X}_n$ as the set of weighted graphs of bounded order n. In addition, we assume that all graphs live on the same set $\mathcal{V} = \{1, \ldots, n\}$ of vertices. Since we are not interested in a specific choice of n, we simply write $\mathcal{X}$ instead of $\mathcal{X}_n$ as the set of all weighted graphs of bounded order n.
Remark 1: It is important to note that specifying an order n and aligning smaller graphs to graphs of order n are purely technical assumptions to simplify mathematics, which can be safely ignored in a practical setting.
We call a weighted graph X*proper* if its edge set is a subset of $\mathcal{R}(X) \times \mathcal{R}(X)$. An example of proper graphs are contact graphs. Suppose that $X = (\mathcal{V}, \mathcal{E}, \omega)$ and $X' = (\mathcal{V}, \mathcal{E}', \omega')$ are weighted graphs living on the common vertex set $\mathcal{V}$. An *r-isomorphism* of X and X' is a bijection $\alpha : \mathcal{V} \to \mathcal{V}$ such that

1) $i < j \Rightarrow \alpha(i) < \alpha(j)$ for all $i, j \in \mathcal{R}(X)$.

2) $(i, j) \in \mathcal{E} \Leftrightarrow (\alpha(i), \alpha(j)) \in \mathcal{E}'$ for all $i, j \in \mathcal{V}$.

In this case, we call X and X' *r-isomorphic*, written as $X \simeq X'$. The set

$$[X] = \{X' \in \mathcal{X} : X \simeq X'\}$$

is called the *r-isomorphism-class* of X. By $[X]$ we denote the set of all r-isomorphism-classes in $\mathcal{X}$. In contrast to the standard definition of isomorphisms in graph theory, an r-isomorphism is order preserving on the residues of X. The notion of r-isomorphism and alignment coincide for proper graphs. It is easy to verify, that the set of r-isomorphisms on X together with the composition forms a subgroup of the permutation group on $\mathcal{V}$.
Let $X = (\mathcal{V}, \mathcal{E}, \omega) \in \mathcal{X}$ be a weighted graph of bounded order n with matrix representation $\boldsymbol{X}$. By stacking the columns of $\boldsymbol{X}$, we can identify X with an N-dimensional vector $\boldsymbol{x}$ from the Euclidean space $\mathbb{E} = \mathbb{R}^N (N = n^2)$. Thus, we have a bijection

$$vec : \mathcal{X} \to \mathbb{E}, \quad X \mapsto vec(X) = \boldsymbol{x}$$

that maps a graph to its *vector representation*. Using $vec(.)$, an r-isomorphic class$[X]$ can be identified with its vector representation $[vec(X)] = \{vec(X') \in \mathbb{E} : X' \in [X]\}$.

By $[\mathbb{E}]$ we denote the set of vector representations of r-isomorphic classes.
An inner product $\langle \cdot, \cdot \rangle$ on $\mathbb{E}$ gives rise to a similarity function on $[\mathbb{E}]$

$$S([\boldsymbol{x}], [(y)]) = max\{\langle \boldsymbol{x'}, \boldsymbol{y'} \rangle : \boldsymbol{x'} \in [\boldsymbol{x}], \boldsymbol{y'} \in [\boldsymbol{y}]\}$$

Suppose that $\boldsymbol{x}$ and $\boldsymbol{y}$ are vector representations of contact graphs X and Y. For the standard inner product, $S(\boldsymbol{x}, \boldsymbol{y})$ is exactly the number of common edges of an optimal pairwise alignment of X and Y. Any inner product space $\mathbb{E}$ is a normed space with norm $\|X\| = \sqrt{\langle \boldsymbol{x}, \boldsymbol{x} \rangle}$. The norm $\| \cdot \|$ on $\mathbb{E}$ induces a function $\ell : [E] \to \mathbb{R}$ with $\ell([\boldsymbol{x}]) = \|\boldsymbol{x}\|$. The function ℓ is independent from the choice of $\boldsymbol{x'} \in [\boldsymbol{x}]$ and therefore well-defined.
Any normed space $\mathbb{E}$ is a metric space with metric $d(\boldsymbol{x}, \boldsymbol{y}) = \|\boldsymbol{x} - \boldsymbol{y}\|$. The metric d on $\mathbb{E}$ induces a disance function on $[\mathbb{E}]$

$$D([\boldsymbol{x}], [\boldsymbol{y}]) = min\{\|\boldsymbol{x'} - \boldsymbol{y'}\| : \boldsymbol{x'} \in [\boldsymbol{x}], \boldsymbol{y'} \in [\boldsymbol{y}]\}$$

Proposition 1: Let $[\boldsymbol{x}], [\boldsymbol{y}] \in [\mathbb{E}]$. Then

$$D([\boldsymbol{x}], [\boldsymbol{y}]^2 = \ell([\boldsymbol{x}])^2 - 2S([\boldsymbol{x}, \boldsymbol{y}]) + \ell([\boldsymbol{y}])^2$$

Proof: Let $\boldsymbol{x'} \in [\boldsymbol{x}]$ and $\boldsymbol{y'} \in [\boldsymbol{y}]$ be vector representations with $D([\boldsymbol{x}], [\boldsymbol{y}]) = \|\boldsymbol{x'} - \boldsymbol{y'}\|$. Then

$$\begin{aligned}
\|\boldsymbol{x'} - \boldsymbol{y'}\|^2 &= \|\boldsymbol{x'}\|^2 - 2\langle \boldsymbol{x'}, \boldsymbol{y'} \rangle + \|\boldsymbol{y'}\|^2 \\
&= \min_{\boldsymbol{x''} \in [\boldsymbol{x}], y'' \in [\boldsymbol{y}]} \|\boldsymbol{x''}\|^2 - 2\langle \boldsymbol{x''}, \boldsymbol{y''} \rangle + \|\boldsymbol{y''}\|^2 \\
&= \|\boldsymbol{x''}\|^2 - 2 \max_{\boldsymbol{x''} \in [\boldsymbol{x}], y'' \in [\boldsymbol{y}]} \langle \boldsymbol{x''}, \boldsymbol{y''} \rangle + \|\boldsymbol{y''}\|^2 \\
&= \ell([\boldsymbol{x}])^2 - 2S([\boldsymbol{x}, [\boldsymbol{y}]]) + \ell([\boldsymbol{y}])^2
\end{aligned}$$

Let $\mathcal{S} = \{X_1, \ldots, X_k\} \subseteq \mathcal{X}$ be a set of k weighted graphs with vector representations $\mathcal{S}_{\mathbb{E}} = \{\boldsymbol{x_1}, \ldots, \boldsymbol{x_k}\} \subseteq \mathbb{E}$. For proper graphs, the set $\mathcal{A}(\mathcal{S})$ of multiple alignments of $\mathcal{S}$. Minimizing the cost function of the structural sample mean is equivalent to minimizing the cost function

$$J_{\mathbb{E}} : \mathbb{E} \to \mathbb{R}, \quad \boldsymbol{x} \mapsto \frac{1}{2} \sum_{p=1}^{k} D([\boldsymbol{x}], [\boldsymbol{x_p}])^2 \tag{A.1}$$

Proposition 2: The function $J_{\mathbb{E}}$ is locally Lipschitz and has a global minimum.

Proof: The function $J_{\mathbb{E}}$ is locally Lipschitz, because the pointwise minimizer of smooth functions is locally Lipschitz[7]. Hence, as a locally Lipschitz function $J_{\mathbb{E}}$ is continuous. Let $c = J_{\mathbb{E}}(\boldsymbol{x_*})$ for some arbitrary $\boldsymbol{x_*} \in \mathbb{E}$, and let $\mathcal{U} = \{\boldsymbol{x} \in \mathbb{E} : J_{\mathbb{E}}(\boldsymbol{x}) \leq c\}$. Since, $J_{\mathbb{E}}$ is continuous, $\mathcal{U}$ is closed. In addition, $\mathcal{U}$ is also bounded. To see this, assume that $\mathcal{U}$ is unbounded. Then there is a sequence $(\boldsymbol{y_i})_{\boldsymbol{i} \in \mathbb{N}}$ in $\mathcal{U}$ with $\lim_{i \to \infty} \|\boldsymbol{y_i}\| = \infty$. From this together with the fact that r-isomorphism classes are finite follows

$$\begin{aligned}
\lim_{i \to \infty} J_{\mathbb{E}}(\boldsymbol{y_i}) &= \lim_{i \to \infty} \sum_{p=1}^{k} \min_{\boldsymbol{x'_p} \in [\boldsymbol{x_p}]} \left\| \boldsymbol{y_i} - \boldsymbol{x'_p} \right\|^2 \\
&\geq \lim_{i \to \infty} \min_{\boldsymbol{x'_1} \in [\boldsymbol{x_1}]} \|\boldsymbol{y_i} - \boldsymbol{x'_1}\|^2 \\
&= \infty
\end{aligned}$$

This contradicts $f(\boldsymbol{y_i}) \leq c$ for all $i \in \mathbb{N}$. Hence, $\mathcal{U}$ is closed and bounded. By the Heine-Borel Theorem, $\mathcal{U}$ is compact. The assertion follows from the fact that a continuous function attains its minimum on a compact set.

Theorem 1: Let $\mathcal{S} = \{X_1, \ldots, X_k\} \in \mathcal{X}$ be a set of k eighted graphs with vector representations $\mathcal{S}_{\mathbb{E}} = \{\boldsymbol{x_1}, \ldots, \boldsymbol{x_k}\} \subseteq \mathbb{E}$. Suppose $\boldsymbol{m} \in \mathbb{E}$ is a global minimum of $J_{\mathbb{E}}$ of Equation (A.1). Then

$$\boldsymbol{m} = \frac{1}{k} \sum_{p=1}^{k} \boldsymbol{x'_p}$$

where $\boldsymbol{x'_p} \in [\boldsymbol{x_p}]$ with $S([\boldsymbol{m}], [\boldsymbol{x_p}]) = \left\langle \boldsymbol{m}, \boldsymbol{x'_p} \right\rangle$ for all $p \in \{1, \ldots, k\}$.
Proof: Let $\mathcal{I} = \{1, \ldots, k\}$ denote the set of indexes. Choose $\boldsymbol{x'_p} \in [\boldsymbol{x_p}]$ with $S([\boldsymbol{m}], [\boldsymbol{x_p}]) = \left\langle \boldsymbol{m}, \boldsymbol{x'_p} \right\rangle$ for all $p \in \mathcal{I}$. Then the function $J_{\mathbb{E}}(\boldsymbol{x}|\boldsymbol{X'}) = \sum_{p \in \mathcal{I}} \left\| \boldsymbol{x} - \boldsymbol{x'_p} \right\|^2$ with $\boldsymbol{X'} = (\boldsymbol{x'_1}, \ldots, \boldsymbol{x'_k}) \in \mathcal{A}(\mathcal{S}_{\mathbb{E}})$ is the optimization formulation of the standard sample mean $\boldsymbol{x_*}$. Then $J(\boldsymbol{x_*}|\boldsymbol{X'}) < J(\boldsymbol{m}|\boldsymbol{X'})$, where strict inequality follows from the fact that the global minimum of $J(\boldsymbol{x}|\boldsymbol{X'})$ is unique. Since $\boldsymbol{m}$ is a global minimum of J by assumption, we find that

$$J(\boldsymbol{x_*}) \geq J(\boldsymbol{m}) > J(\boldsymbol{x_*}|\boldsymbol{X'}) \tag{A.2}$$

Since $J(\boldsymbol{x_*}) \neq J(\boldsymbol{x_*}|\boldsymbol{X'})$ there is a nonempty subset $\mathcal{J} \subseteq \mathcal{I}$ with $S([\boldsymbol{x_*}], [\boldsymbol{x_q}]) > \left\langle \boldsymbol{x_*}, \boldsymbol{x'_q} \right\rangle$ for all $q \in \mathcal{J}$. Let $\tilde{\boldsymbol{x}}_{\boldsymbol{q}} \in [\boldsymbol{x_q}]$ with $S([\boldsymbol{x_*}], [\boldsymbol{x_q}]) = \langle (x_*), \tilde{\boldsymbol{x}}_{\boldsymbol{q}} \rangle$ for all $q \in \mathcal{J}$. From $D([\boldsymbol{x_*}], [\tilde{\boldsymbol{x}}_{\boldsymbol{q}}]) = \|\boldsymbol{x_*} - \tilde{\boldsymbol{x}}_{\boldsymbol{q}}\| < \left\| \boldsymbol{x_*} - \boldsymbol{x'_q} \right\|$ for all $q \in \mathcal{J}$ follows

$$\begin{aligned} J(\boldsymbol{x_*}) &= \sum_{p \in \mathcal{I} \setminus \mathcal{J}} \left\| \boldsymbol{x_*} - \boldsymbol{x'_p} \right\|^2 + \sum_{q \in \mathcal{J}} \left\| \boldsymbol{x_*} - \tilde{\boldsymbol{x}}_{\boldsymbol{q}} \right\|^2 \\ &= \sum_{p \in \mathcal{I} \setminus \mathcal{J}} \left\| \boldsymbol{x_*} - \boldsymbol{x'_p} \right\|^2 + \sum_{q \in \mathcal{I}} \left\| \boldsymbol{x_*} - \boldsymbol{x'_q} \right\|^2 \\ &= J(\boldsymbol{x_*}|\boldsymbol{X'}) \end{aligned}$$

Inequality $J(\boldsymbol{x_*}) < J(\boldsymbol{x_*}|\boldsymbol{X'})$ cotradicts Equation A.2. Hence we find that $\boldsymbol{m'} = \boldsymbol{x_*}$. The *Gram sum* of $\boldsymbol{X} = (\boldsymbol{x_1}, \ldots, \boldsymbol{x_k}) \in \mathcal{A}(\mathcal{S}_{\mathbb{E}})$ is defined by $G(\boldsymbol{X}) = \sum_p \sum_{q>p} \langle \boldsymbol{x_p}, \boldsymbol{x_q} \rangle$.

Theorem 2: Let $S = \{X_1, \ldots, X_k\} \in \mathcal{X}$ be a set of k weighted graphs. .Suppose that $\boldsymbol{X} = (\boldsymbol{x_1}, \ldots, \boldsymbol{x_k}) \in \mathcal{A}(\mathcal{S}_{\mathbb{E}})$. Then the following statements are equivalent:

1) The sample mean $\boldsymbol{m}$ of $\boldsymbol{x_1}, \ldots, \boldsymbol{x_k}$ is a global minimum of $J_{\mathbb{E}}$.

2) $G(\boldsymbol{X}) \geq G(\boldsymbol{X'})$ for all $\boldsymbol{X'} \in \mathcal{A}(\mathcal{S}_{\mathbb{E}})$.

Proof: It is sufficient to show that

$$J_{\mathbb{E}}(\boldsymbol{m}) = \sum_{p=1}^{k} \|\boldsymbol{x_p}\|^2 - \frac{1}{k} G(\boldsymbol{X}) = \sum_{p=1}^{k} \ell(\boldsymbol{x_p})^2 - \frac{1}{k} G(\boldsymbol{X})$$

We have

$$
\begin{aligned}
J_{\mathbb{E}}(\boldsymbol{m}) &= \sum_{p=1}^{k} \|\boldsymbol{m} - \boldsymbol{x_p}\|^2 \\
&= \sum_{p=1}^{k} \|\boldsymbol{m}\|^2 - 2\langle \boldsymbol{m}, \boldsymbol{x_p}\rangle + \|\boldsymbol{x_p}\|^2 \\
&= \sum_{p=1}^{k} \left\{ \left\| \frac{1}{k} \sum_{q=1}^{k} \boldsymbol{x_q} \right\|^2 - 2\left\langle \frac{1}{k} \sum_{q=1}^{k} \boldsymbol{x_q}, \boldsymbol{x_p} \right\rangle + \|\boldsymbol{x_p}\|^2 \right\} \\
&= \sum_{p=1}^{k} \left\{ \frac{1}{k^2} \sum_{r=1}^{k} \sum_{s=1}^{k} \langle \boldsymbol{x_r}, \boldsymbol{x_s}\rangle - \frac{2}{k} \sum_{q=1}^{k} \langle \boldsymbol{x_q}, \boldsymbol{x_p}\rangle + \|\boldsymbol{x_p}\|^2 \right\} \\
&= \frac{1}{k} \sum_{p=1}^{k} \sum_{q=1}^{k} \langle \boldsymbol{x_p}, \boldsymbol{x_q}\rangle - \frac{2}{k} \sum_{p=1}^{k} \sum q = 1^k \langle \boldsymbol{x_p}, \boldsymbol{x_q}\rangle + \sum_{p=1}^{k} \|\boldsymbol{x_p}\|^2 \\
&= \sum p = 1^k \|\boldsymbol{x_p}\|^2 - \frac{1}{k} \underbrace{\sum_{p=1}^{k} \sum_{q=1}^{k} \langle \boldsymbol{x_p}, \boldsymbol{x_q}\rangle}_{=G(\boldsymbol{X})}
\end{aligned}
$$

Appendix B

Zusammenfassung

Die dreidimensionale Struktur eines Proteins wird bestimmt durch die kovalenten und nicht-kovalenten Wechselwirkungen seiner Aminosäuren. Die genaue Beschreibung dieser Wechselwirkungen erfordert die Berücksichtigung von Quanteneffekten, was das System zu komplex für viele Analysen macht. Ja nach Anwendung kann ein geeigneteres Modell gewählt werden, das die Komplexität reduziert und gleichzeitig die für die Anwendung relevanten Eigenschaften erhält. In dieser Arbeit erforschen wir die Darstellung von Proteinen als Netzwerk von interagierenden Aminosäuren. Diese Graphdarstellung hat den Vorteil, dass sie die Anwendung von Methoden aus der Graphentheorie erlaubt und gleichzeitig die Information über die Tertiärstruktur erhält. Auf der Grundlage dieser Darstellung entwickeln wir neue Methoden für drei wichtige Probleme der Strukturbiologie: Die Vorhersage der Struktur aus der Sequenz, der Vergleich von Strukturen und die Auswirkungen von Mutationen auf die Proteinstruktur und -funktion. In Teil 1 zeigen wir, wie eine graphentheoretisch motivierte Definition des multiplen Strukturalignmentproblems auf eine überraschende Parallele zur Theorie des arithmetischen Mittels von Graphen führt. Wir zeigen, dass die Berechnung des Graph-Mittels äquivalent zur Berechnung des optimalen Struktur-Alignment ist und wie dies zu einer neuen multiplen Strukturalignmentmethode führt. In Teil 2 verwenden wir die Graph-Mittel-Theorie, um eine Methode zur konsensbasierten Vorhersage von Intraproteinkontakten zu entwicklen. Wir zeigen, dass mit Hilfe dieser Methode, Kontake innerhalb von Proteinen besser vorhergesagt werden können, als mit jeder anderen aktuell verfügbaren Methode. In Teil 3 zeigen wir, wie Strukturinformationen genutzt werden können, um die Auswirkungen von Krebsmutationen funktionell zu analysieren. Für fast 2000 Mutationen analysieren wir deren Auswirkung auf die jeweilge Proteinstruktur. Wir zeigen, dass die Mutationsmuster sich stark unterscheiden zwischen Mutationen in Onkogenen und Mutationen in Tumorsuppressoren. Dies führt zu einer Methode, mit deren Hilfe sich anhand der Mutationsmuster prediktive Aussagen machen lassen, ob ein unbekanntes Protein sich im Krebs-Zusammenhang wie ein Onkogen oder ein Tumorsupporessor verhält. Wir spannen mit dieser Arbeit den Bogen von Sequenzinformation über Struktur hin zur funktionellen Analyse. Je mehr sich der aktuelle Trend der immer schnelleren Generierung von Sequenzdaten fortsetzt, desto mehr werden effiziente Methoden zur strukturellen und funktionellen Analyse an Bedeutung gewinnen.

Bibliography

Alber, F., Dokudovskaya, S., Veenhoff, L. M., Zhang, W., Kipper, J., Devos, D., Suprapto, A., Karni-Schmidt, O., Williams, R., Chait, B. T., Rout, M. P., and Sali, A. (2007). Determining the architectures of macromolecular assemblies. *Nature*, 450(7170):683–694.

Altschuh, D., Lesk, A. M., Bloomer, A. C., and Klug, A. (1987). Correlation of co-ordinated amino acid substitutions with function in viruses related to tobacco mosaic virus. *J Mol Biol*, 193(4):693–707.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J Mol Biol*, 215(3):403–410.

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997). Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic Acids Res*, 25(17):3389–3402.

Amitai, G., Shemesh, A., Sitbon, E., Shklar, M., Netanely, D., Venger, I., and Pietrokovski, S. (2004). Network analysis of protein structures identifies functional residues. *J Mol Biol*, 344(4):1135–1146.

Anfinsen, C. B. (1973). Principles that govern the folding of protein chains. *Science*, 181(96):223–230.

Anson, M. L. and Mirsky, A. E. (1925). On some general properties of proteins. *J Gen Physiol*, 9(2):169–179.

Baker, D. and Sali, A. (2001). Protein structure prediction and structural genomics. *Science*, 294(5540):93–96.

Ben-David, M., Noivirt-Brik, O., Paz, A., Prilusky, J., Sussman, J. L., and Levy, Y. (2009). Assessment of Casp8 structure predictions for template free targets. *Proteins*, 77 Suppl 9:50–65.

Berman, H., Henrick, K., and Nakamura, H. (2003). Announcing the worldwide protein data bank. *Nat Struct Biol*, 10(12):980.

Berman, H. M. (2008). The protein data bank: a historical perspective. *Acta Crystallogr A*, 64(Pt 1):88–95.

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

Bozic, I., Antal, T., Ohtsuki, H., Carter, H., Kim, D., Chen, S., Karchin, R., Kinzler, K. W., Vogelstein, B., and Nowak, M. A. (2010). Accumulation of driver and passenger mutations during tumor progression. *Proc Natl Acad Sci U S A*, 107(43):18545–50.

Browne, W. J., North, A. C., Phillips, D. C., Brew, K., Vanaman, T. C., and Hill, R. L. (1969). A possible three-dimensional structure of bovine alpha-lactalbumin based on that of hen's egg-white lysozyme. *J Mol Biol*, 42(1):65–86.

Buchete, N.-V., Straub, J. E., and Thirumalai, D. (2004). Orientational potentials extracted from protein structures improve native fold recognition. *Protein Sci*, 13(4):862–874.

Bujnicki, J. M., Elofsson, A., Fischer, D., and Rychlewski, L. (2001). Structure prediction meta server. *Bioinformatics*, 17(8):750–751.

Burkert, U. (1982). *Molecular mechanics*. American Chemical Society, Washington, D.C.

Cancer Genome Atlas Research Network (2008). Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, 455(18772890):1061–8.

Caprara, A., Carr, R., Istrail, S., Lancia, G., and Walenz, B. (2004). 1001 optimal pdb structure alignments: integer programming methods for finding the maximum contact map overlap. *J Comput Biol*, 11(1):27–52.

Caprara, A. and Lancia, G. (2002). Optimal and near-optimal solutions for 3d structure comparisons. In *Proc. First Int 3D Data Processing Visualization and Transmission Symp*, pages 737–744.

Carter, H., Chen, S., Isik, L., Tyekucheva, S., Velculescu, V. E., Kinzler, K. W., Vogelstein, B., and Karchin, R. (2009). Cancer-specific high-throughput annotation of somatic mutations: computational prediction of driver missense mutations. *Cancer Research*, 69(19654296):6660–7.

Cheng, J. and Baldi, P. (2007). Improved residue contact prediction using support vector machines and a large feature set. *BMC Bioinformatics*, 8:113.

Chick, H. and Martin, C. (1910). On the "heat" coagulation of proteins. *Journal of Physiology*, 40:404–—430.

Chittenden, T. W., Howe, E. A., Culhane, A. C., Sultana, R., Taylor, J. M., Holmes, C., and Quackenbush, J. (2008). Functional classification analysis of somatically mutated genes in human breast and colorectal cancers. *Genomics*, 91(6):508–11.

Chothia, C. and Lesk, A. M. (1986). The relation between the divergence of sequence and structure in proteins. *EMBO J*, 5(4):823–826.

Cline, M. S., Karplus, K., Lathrop, R. H., Smith, T. F., Rogers, R. G., and Haussler, D. (2002). Information-theoretic dissection of pairwise contact potentials. *Proteins*, 49(1):7–14.

Cormen, T. H., Leiserson, C. E., and Rivest, R. L. (2001). *Introduction to Algorithms (2nd edition)*. MIT Press and McGraw–Hill.

Crippen, G. and Havel, T. (1988). *Distance Geometry and Molecular Conformation*. John Wiley & Sons.

Cuff, J. A., Clamp, M. E., Siddiqui, A. S., Finlay, M., and Barton, G. J. (1998). Jpred: a consensus secondary structure prediction server. *Bioinformatics*, 14(10):892–893.

Davies, H., Bignell, G. R., Cox, C., Stephens, P., Edkins, S., Clegg, S., Teague, J., Woffendin, H., Garnett, M. J., Bottomley, W., Davis, N., Dicks, E., Ewing, R., Floyd, Y., Gray, K., Hall, S., Hawes, R., Hughes, J., Kosmidou, V., Menzies, A., Mould, C., Parker, A., Stevens, C., Watt, S., Hooper, S., Wilson, R., Jayatilake, H., Gusterson, B. A., Cooper, C., Shipley, J., Hargrave, D., Pritchard-Jones, K., Maitland, N., Chenevix-Trench, G., Riggins, G. J., Bigner, D. D., Palmieri, G., Cossu, A., Flanagan, A., Nicholson, A., Ho, J. W., Leung, S. Y., Yuen, S. T., Weber, B. L., Seigler, H. F., Darrow, T. L., Paterson, H., Marais, R., Marshall, C. J., Wooster, R., Stratton, M. R., and Futreal, P. A. (2002). Mutations of the braf gene in human cancer. *Nature*, 417(6892):949–54.

de Berg, M., Cheong, O., van Kreveld, M., and Overmars, M. (2008). *Computational geometry : algorithms and applications*. Springer, Berlin.

del Sol, A., Fujihashi, H., and O'Meara, P. (2005). Topology of small-world networks of protein-protein complex structures. *Bioinformatics*, 21(8):1311–1315.

Desmet, J., Maeyer, M. D., Hazes, B., and Lasters, I. (1992). The dead-end elimination theorem and its use in protein side-chain positioning. *Nature*, 356(6369):539–542.

Dill, K. A. and Chan, H. S. (1997). From levinthal to pathways to funnels. *Nat Struct Biol*, 4(1):10–19.

Duan, Y. and Kollman, P. A. (1998). Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution. *Science*, 282(5389):740–744.

Duarte, J. M., Sathyapriya, R., Stehr, H., Filippis, I., and Lappe, M. (2010). Optimal contact definition for reconstruction of contact maps. *BMC Bioinformatics*, 11:283.

Dunbrack, R. L. and Karplus, M. (1993). Backbone-dependent rotamer library for proteins. application to side-chain prediction. *J Mol Biol*, 230(2):543–574.

Edsall, J. T. (1995). Hsien wu and the first theory of protein denaturation (1931). *Adv Protein Chem*, 46:1–5.

Ezkurdia, I., na, O. G., Izarzugaza, J. M. G., and Tress, M. L. (2009). Assessment of domain boundary predictions and the prediction of intramolecular contacts in casp8. *Proteins*, 77 Suppl 9:196–209.

Fariselli, P., Olmea, O., Valencia, A., and Casadio, R. (2001). Progress in predicting inter-residue contacts of proteins with neural networks and correlated mutations. *Proteins*, Suppl 5:157–162.

Fiser, A., Do, R. K., and Sali, A. (2000). Modeling of loops in protein structures. *Protein Sci*, 9(9):1753–1773.

Fiser, A. and Sali, A. (2003). Modloop: automated modeling of loops in protein structures. *Bioinformatics*, 19(18):2500–2501.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7:179–188.

Flöckner, H., Braxenthaler, M., Lackner, P., Jaritz, M., Ortner, M., and Sippl, M. J. (1995). Progress in fold recognition. *Proteins*, 23(3):376–386.

Fodor, A. A. and Aldrich, R. W. (2004). Influence of conservation on calculations of amino acid covariance in multiple sequence alignments. *Proteins*, 56(2):211–221.

Forbes, S. A., Bhamra, G., Bamford, S., Dawson, E., Kok, C., Clements, J., Menzies, A., Teague, J. W., Futreal, P. A., and Stratton, M. R. (2008). The catalogue of somatic mutations in cancer (cosmic). *Curr Protoc Hum Genet*, Chapter 10:Unit 10 11.

Freudenthal, H. E. (1961). The concept and the role of the model in mathematics and natural and social sciences. In *Proceedings of the Colloquium sponsored by the Division of Philosophy of Sciences of the International Union of History and Philosophy of Sciences organized at Utrecht, January 1960*.

Förster, F., Lasker, K., Nickell, S., Sali, A., and Baumeister, W. (2010). Toward an integrated structural model of the 26s proteasome. *Mol Cell Proteomics*, 9(8):1666–1677.

Fujitsuka, Y., Chikenji, G., and Takada, S. (2006). Simfold energy function for de novo protein structure prediction: consensus with rosetta. *Proteins*, 62(2):381–398.

Garey, M. R. and Johnson, D. S. (1979). *Computers and Intractability: A Guide to the Theory of NP-Completeness*. W. H. Freeman & Co., New York, NY, USA.

Gerstein, M. and Levitt, M. (1996). Using iterative dynamic programming to obtain accurate pairwise and multiple alignments of protein structures. *Proc Int Conf Intell Syst Mol Biol*, 4:59–67.

Göbel, U., Sander, C., Schneider, R., and Valencia, A. (1994). Correlated mutations and residue contacts in proteins. *Proteins*, 18(4):309–317.

Godzik, A., Kolinski, A., and Skolnick, J. (1992). Topology fingerprint approach to the inverse protein folding problem. *J Mol Biol*, 227(1):227–238.

Godzik, A. and Skolnick, J. (1994). Flexible algorithm for direct multiple alignment of protein structures and sequences. *Comput Appl Biosci*, 10(6):587–596.

Goldman, D., Istrail, S., and Papadimitriou, C. H. (1999). Algorithmic aspects of protein structure similarity. In *Proc. 40th Annual Symp. Foundations of Computer Science*, pages 512–521.

Grana, O., Baker, D., MacCallum, R. M., Meiler, J., Punta, M., Rost, B., Tress, M. L.,

and Valencia, A. (2005). Casp6 assessment of contact prediction. *Proteins*, 61 Suppl 7:214–224.

Greenman, C., Stephens, P., Smith, R., Dalgliesh, G. L., Hunter, C., Bignell, G., Davies, H., Teague, J., Butler, A., Stevens, C., Edkins, S., O'Meara, S., Vastrik, I., Schmidt, E. E., Avis, T., Barthorpe, S., Bhamra, G., Buck, G., Choudhury, B., Clements, J., Cole, J., Dicks, E., Forbes, S., Gray, K., Halliday, K., Harrison, R., Hills, K., Hinton, J., Jenkinson, A., Jones, D., Menzies, A., Mironenko, T., Perry, J., Raine, K., Richardson, D., Shepherd, R., Small, A., Tofts, C., Varian, J., Webb, T., West, S., Widaa, S., Yates, A., Cahill, D. P., Louis, D. N., Goldstraw, P., Nicholson, A. G., Brasseur, F., Looijenga, L., Weber, B. L., Chiew, Y. E., DeFazio, A., Greaves, M. F., Green, A. R., Campbell, P., Birney, E., Easton, D. F., Chenevix-Trench, G., Tan, M. H., Khoo, S. K., Teh, B. T., Yuen, S. T., Leung, S. Y., Wooster, R., Futreal, P. A., and Stratton, M. R. (2007). Patterns of somatic mutation in human cancer genomes. *Nature*, 446(7132):153–8.

Guda, C., Scheeff, E. D., Bourne, P. E., and Shindyalov, I. N. (2001). A new algorithm for the alignment of multiple protein structures using monte carlo optimization. *Pac Symp Biocomput*, pages 275–286.

Guerler, A. and Knapp, E.-W. (2010). Strategies of non-sequential protein structure alignments. *Genome Inform*, 22:21–29.

Guerois, R., Nielsen, J. E., and Serrano, L. (2002). Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. *J Mol Biol*, 320(2):369–87.

Hamilton, N., Burrage, K., Ragan, M. A., and Huber, T. (2004). Protein contact prediction using patterns of correlation. *Proteins*, 56(4):679–684.

Hamilton, W. (1971). Protein data bank. *Nature New Biol*, 233:223.

Hildebrand, A., Remmert, M., Biegert, A., and Söding, J. (2009a). Fast and accurate automatic structure prediction with hhpred. *Proteins*, 77 Suppl 9:128–132.

Hildebrand, P. W., Goede, A., Bauer, R. A., Gruening, B., Ismer, J., Michalsky, E., and Preissner, R. (2009b). Superlooper–a prediction server for the modeling of loops in globular and membrane proteins. *Nucleic Acids Res*, 37(Web Server issue):W571–W574.

Hirschman, L., Yeh, A., Blaschke, C., and Valencia, A. (2005). Overview of biocreative: critical assessment of information extraction for biology. *BMC Bioinformatics*, 6 Suppl 1:S1.

Holm, L. and Sander, C. (1991). Database algorithm for generating protein backbone and side-chain co-ordinates from a c alpha trace application to model building and detection of co-ordinate errors. *J Mol Biol*, 218(1):183–194.

Holm, L. and Sander, C. (1993). Protein structure comparison by alignment of distance matrices. *J Mol Biol*, 233(1):123–138.

Hornbeck, P. V., Chabra, I., Kornhauser, J. M., Skrzypek, E., and Zhang, B. (2004).

Phosphosite: A bioinformatics resource dedicated to physiological protein phosphorylation. *Proteomics*, 4(6):1551–61.

Huan, J., Bandyopadhyay, D., Wang, W., Snoeyink, J., Prins, J., and Tropsha, A. (2005). Comparing graph representations of protein structure for mining family-specific residue-based packing motifs. *J Comput Biol*, 12(6):657–671.

Huang, C. H., Mandelker, D., Gabelli, S. B., and Amzel, L. M. (2008). Insights into the oncogenic effects of pik3ca mutations from the structure of p110alpha/p85alpha. *Cell Cycle*, 7(9):1151–6.

Hubbard, S. J. and Thornton, J. M. (1993). Naccess, computer program, department of biochemistry and molecular biology, university college london.

Hurst, J. M., McMillan, L. E., Porter, C. T., Allen, J., Fakorede, A., and Martin, A. C. (2009). The saapdb web resource: a large-scale structural analysis of mutant proteins. *Hum Mutat*, 30(4):616–24.

Ilinkin, I., Ye, J., and Janardan, R. (2010). Multiple structure alignment and consensus identification for proteins. *BMC Bioinformatics*, 11:71.

Istrail, S. and Lam, F. (2009). Combinatorial algorithms for protein folding in lattice models: A survey of mathematical results. *Communications in Information and Systems*, 9(4):303–346.

Izarzugaza, J. M. G., na, O. G., Tress, M. L., Valencia, A., and Clarke, N. D. (2007). Assessment of intramolecular contact predictions for Casp7. *Proteins*, 69 Suppl 8:152–158.

Jacobson, M. P., Pincus, D. L., Rapp, C. S., Day, T. J. F., Honig, B., Shaw, D. E., and Friesner, R. A. (2004). A hierarchical approach to all-atom protein loop prediction. *Proteins*, 55(2):351–367.

Jain, B. and Obermayer, K. (2008). On the sample mean of graphs. In *Proc. (IEEE World Congress Computational Intelligence). IEEE Int. Joint Conf. Neural Networks IJCNN 2008*, pages 993–1000.

Jain, B. J. and Obermayer, K. (2009a). Algorithms for the sample mean of graphs. In *Proceedings of the 13th International Conference on Computer Analysis of Images and Patterns*, CAIP '09, pages 351–359, Berlin, Heidelberg. Springer-Verlag.

Jain, B. J. and Obermayer, K. (2009b). Bimal: Bipartite matching alignment for the contact map overlap problem. In *Proc. Int. Joint Conf. Neural Networks IJCNN 2009*, pages 1394–1400.

Janin, J., Henrick, K., Moult, J., Eyck, L. T., Sternberg, M. J. E., Vajda, S., Vakser, I., Wodak, S. J., and of PRedicted Interactions, C. A. (2003). Capri: a critical assessment of predicted interactions. *Proteins*, 52(1):2–9.

Jauch, R., Yeo, H. C., Kolatkar, P. R., and Clarke, N. D. (2007). Assessment of Casp7 structure predictions for template free targets. *Proteins*, 69 Suppl 8:57–67.

Jeffers, M., Schmidt, L., Nakaigawa, N., Webb, C. P., Weirich, G., Kishida, T., Zbar,

B., and Vande Woude, G. F. (1997). Activating mutations for the met tyrosine kinase receptor in human cancer. *Proc Natl Acad Sci U S A*, 94(21):11445–50.

Jiang, L., Althoff, E. A., Clemente, F. R., Doyle, L., Röthlisberger, D., Zanghellini, A., Gallaher, J. L., Betker, J. L., Tanaka, F., Barbas, C. F., Hilvert, D., Houk, K. N., Stoddard, B. L., and Baker, D. (2008). De novo computational design of retro-aldol enzymes. *Science*, 319(5868):1387–1391.

Jones, D. T. (1997). Successful ab initio prediction of the tertiary structure of nk-lysin using multiple sequences and recognized supersecondary structural motifs. *Proteins*, Suppl 1:185–191.

Jones, D. T., Bryson, K., Coleman, A., McGuffin, L. J., Sadowski, M. I., Sodhi, J. S., and Ward, J. J. (2005). Prediction of novel and analogous folds using fragment assembly and fold recognition. *Proteins*, 61 Suppl 7:143–151.

Jones, D. T., Taylor, W. R., and Thornton, J. M. (1992). A new approach to protein fold recognition. *Nature*, 358(6381):86–89.

Jones, S., Zhang, X., Parsons, D. W., Lin, J. C., Leary, R. J., Angenendt, P., Mankoo, P., Carter, H., Kamiyama, H., Jimeno, A., Hong, S. M., Fu, B., Lin, M. T., Calhoun, E. S., Kamiyama, M., Walter, K., Nikolskaya, T., Nikolsky, Y., Hartigan, J., Smith, D. R., Hidalgo, M., Leach, S. D., Klein, A. P., Jaffee, E. M., Goggins, M., Maitra, A., Iacobuzio-Donahue, C., Eshleman, J. R., Kern, S. E., Hruban, R. H., Karchin, R., Papadopoulos, N., Parmigiani, G., Vogelstein, B., Velculescu, V. E., and Kinzler, K. W. (2008). Core signaling pathways in human pancreatic cancers revealed by global genomic analyses. *Science (New York, NY)*, 321(18772397):1801–6.

Jones, T. A. and Thirup, S. (1986). Using known substructures in protein model building and crystallography. *EMBO J*, 5(4):819–822.

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

Kaminker, J. S., Zhang, Y., Waugh, A., Haverty, P. M., Peters, B., Sebisanovic, D., Stinson, J., Forrest, W. F., Bazan, J. F., Seshagiri, S., and Zhang, Z. (2007). Distinguishing cancer-associated missense mutations from common polymorphisms. *Cancer Res*, 67(2):465–73.

Karchin, R. (2009). Next generation tools for the annotation of human snps. *Brief Bioinform*, 10(1):35–52.

Karpen, M. E., de Haseth, P. L., and Neet, K. E. (1989). Comparing short protein substructures by a method based on backbone torsion angles. *Proteins*, 6(2):155–167.

Kendrew, J. C., Dickerson, R. E., Strandberg, B. E., Hart, R. G., Davies, D. R., Phillips, D. C., and Shore, V. C. (1960). Structure of myoglobin: A three-dimensional fourier synthesis at 2 a. resolution. *Nature*, 185(4711):422–427.

Khersonsky, O., Röthlisberger, D., Wollacott, A. M., Murphy, P., Dym, O., Albeck, S., Kiss, G., Houk, K. N., Baker, D., and Tawfik, D. S. (2011). Optimization of

the in-silico-designed kemp eliminase ke70 by computational design and directed evolution. *J Mol Biol*, 407(3):391–412.

Kirkpatrick, S., Gelatt, C. D., and Vecchi, M. P. (1983). Optimization by simulated annealing. *Science*, 220(4598):671–680.

Klammer, M., Messina, D. N., Schmitt, T., and Sonnhammer, E. L. L. (2009). Metatm - a consensus method for transmembrane protein topology prediction. *BMC Bioinformatics*, 10:314.

Kolodny, R., Koehl, P., Guibas, L., and Levitt, M. (2002). Small libraries of protein fragments model native protein structures accurately. *J Mol Biol*, 323(2):297–307.

Kolodny, R., Koehl, P., and Levitt, M. (2005). Comprehensive evaluation of protein structure alignment methods: scoring by geometric measures. *J Mol Biol*, 346(4):1173–1188.

Konagurthu, A. S., Whisstock, J. C., Stuckey, P. J., and Lesk, A. M. (2006). Mustang: a multiple structural alignment algorithm. *Proteins*, 64(3):559–574.

Krissinel, E. and Henrick, K. (2004). Secondary-structure matching (ssm), a new tool for fast protein structure alignment in three dimensions. *Acta Crystallogr D Biol Crystallogr*, 60(Pt 12 Pt 1):2256–2268.

Kuhlman, B., Dantas, G., Ireton, G. C., Varani, G., Stoddard, B. L., and Baker, D. (2003). Design of a novel globular protein fold with atomic-level accuracy. *Science*, 302(5649):1364–1368.

Latek, D. and Kolinski, A. (2008). Contact prediction in protein modeling: scoring, folding and refinement of coarse-grained models. *BMC Struct Biol*, 8:36.

Lathrop, R. H. (1994). The protein threading problem with sequence amino acid interaction preferences is np-complete. *Protein Eng*, 7(9):1059–1068.

Lathrop, R. H. and Smith, T. F. (1994). A branch-and-bound algorithm for optimal protein threading with pairwise (contact potential) amino acid interactions. In *Proc. Twenty-Seventh Hawaii Int System Sciences Conf*, volume 5, pages 365–374.

Leach, A. R. (2001). *Molecular modelling : principles and applications*. Prentice Hall.

Lee, D. J., Schonleben, F., Banuchi, V. E., Qiu, W., Close, L. G., Assaad, A. M., and Su, G. H. (2010). Multiple tumor-suppressor genes on chromosome 3p contribute to head and neck squamous cell carcinoma tumorigenesis. *Cancer Biol Ther*, 10(7).

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

Levinthal, C. (1966). Molecular model-building by computer. *Sci Am*, 214(6):42–52.

Levinthal, C. (1969). How to fold graciously. In DeBrunner, J. T. P. and Munck, E., editors, *Mossbauer Spectroscopy in Biological Systems: Proceedings of a meeting held at Allerton House, Monticello, Illinois*. University of Illinois Press. 22–24.

Levitt, M. (1976). A simplified representation of protein conformations for rapid simulation of protein folding. *J Mol Biol*, 104(1):59–107.

Levitt, M. and Sharon, R. (1988). Accurate simulation of protein dynamics in solution. *Proc Natl Acad Sci U S A*, 85(20):7557–7561.

Levitt, M. and Warshel, A. (1975). Computer simulation of protein folding. *Nature*, 253(5494):694–698.

Lund, O., Frimand, K., Gorodkin, J., Bohr, H., Bohr, J., Hansen, J., and Brunak, S. (1997). Protein distance constraints predicted by neural networks and probability density functions. *Protein Eng*, 10(11):1241–1248.

Lupyan, D., Leo-Macias, A., and Ortiz, A. R. (2005). A new progressive-iterative algorithm for multiple structure alignment. *Bioinformatics*, 21(15):3255–3263.

Mäkelä, M. M. and Neittaanmäki, P. (1992). *Nonsmooth Optimization*. World Scientific.

Malumbres, M. and Barbacid, M. (2003). Ras oncogenes: the first 30 years. *Nat Rev Cancer*, 3(6):459–65.

Mandell, D. J., Coutsias, E. A., and Kortemme, T. (2009). Sub-angstrom accuracy in protein loop reconstruction by robotics-inspired conformational sampling. *Nat Methods*, 6(8):551–552.

Margraf, T., Schenk, G., and Torda, A. E. (2009). The salami protein structure search server. *Nucleic Acids Res*, 37(Web Server issue):W480–W484.

May, A. C. and Johnson, M. S. (1995). Improved genetic algorithm-based protein structure comparisons: pairwise and multiple superpositions. *Protein Eng*, 8(9):873–882.

Menke, M., Berger, B., and Cowen, L. (2008). Matt: local flexibility aids protein multiple structure alignment. *PLoS Comput Biol*, 4(1):e10.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state calculations by fast computing machines. *Journal of Chemical Physics*, 21:1087–1092.

Michalsky, E., Goede, A., and Preissner, R. (2003). Loops in proteins (lip)–a comprehensive loop database for homology modelling. *Protein Eng*, 16(12):979–985.

Miller, C. S. and Eisenberg, D. (2008). Using inferred residue contacts to distinguish between correct and incorrect protein models. *Bioinformatics*, 24(14):1575–1582.

Mirsky, A. E. and Pauling, L. (1936). On the structure of native, denatured, and coagulated proteins. *Proc Natl Acad Sci U S A*, 22(7):439–447.

Miyazawa, S. and Jernigan, R. L. (1985). Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. *Macromolecules*, 18(3):534–552.

Miyazawa, S. and Jernigan, R. L. (2005). How effective for fold recognition is a potential of mean force that includes relative orientations between contacting residues in proteins? *J Chem Phys*, 122(2):024901.

Mizuguchi, K., Deane, C. M., Blundell, T. L., and Overington, J. P. (1998). Homstrad: a database of protein structure alignments for homologous families. *Protein Sci*, 7(11):2469–2471.

Moult, J. (2005). A decade of casp: progress, bottlenecks and prognosis in protein structure prediction. *Curr Opin Struct Biol*, 15(3):285–289.

Moult, J. (2006). Rigorous performance evaluation in protein structure modelling and implications for computational biology. *Philos Trans R Soc Lond B Biol Sci*, 361(1467):453–458.

Moult, J. and James, M. N. (1986). An algorithm for determining the conformation of polypeptide segments in proteins by systematic search. *Proteins*, 1(2):146–163.

Neher, E. (1994). How frequent are correlated changes in families of protein sequences? *Proc Natl Acad Sci U S A*, 91(1):98–102.

Ng, P. C. and Henikoff, S. (2003). Sift: Predicting amino acid changes that affect protein function. *Nucleic acids research*, 31(12824425):3812–4.

Ng, P. C. and Henikoff, S. (2006). Predicting the effects of amino acid substitutions on protein function. *Annu Rev Genomics Hum Genet*, 7:61–80.

Nishikawa, K. and Matsuo, Y. (1993). Development of pseudoenergy potentials for assessing protein 3-d-1-d compatibility and detecting weak homologies. *Protein Eng*, 6(8):811–820.

Olmea, O. and Valencia, A. (1997). Improving contact predictions by the combination of correlated mutations and other sources of sequence information. *Fold Des*, 2(3):S25–S32.

O'Sullivan, O., Suhre, K., Abergel, C., Higgins, D. G., and Notredame, C. (2004). 3dcoffee: combining protein sequences and structures within multiple sequence alignments. *J Mol Biol*, 340(2):385–395.

Pai, E. F., Kabsch, W., Krengel, U., Holmes, K. C., John, J., and Wittinghofer, A. (1989). Structure of the guanine-nucleotide-binding domain of the ha-ras oncogene product p21 in the triphosphate conformation. *Nature*, 341(6239):209–14.

Park, J., Karplus, K., Barrett, C., Hughey, R., Haussler, D., Hubbard, T., and Chothia, C. (1998). Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods. *J Mol Biol*, 284(4):1201–1210.

Paszkiewicz, K. H., Sternberg, M. J. E., and Lappe, M. (2006). Prediction of viable circular permutants using a graph theoretic approach. *Bioinformatics*, 22(11):1353–1358.

Pauling, L., Corey, R. B., and Branson, H. R. (1951). The structure of proteins; two

hydrogen-bonded helical configurations of the polypeptide chain. *Proc Natl Acad Sci U S A*, 37(4):205–211.

Perutz, M. (1997). *Science is not a quiet life: unravelling the atomic mechanism of haemoglobin*. World Scientific.

Perutz, M. F., Rossmann, M. G., Cullis, A. F., Muirhead, H., Will, G., and North, A. C. (1960). Structure of haemoglobin: a three-dimensional fourier synthesis at 5.5-a. resolution, obtained by x-ray analysis. *Nature*, 185(4711):416–422.

Petrey, D., Xiang, Z., Tang, C. L., Xie, L., Gimpelev, M., Mitros, T., Soto, C. S., Goldsmith-Fischman, S., Kernytsky, A., Schlessinger, A., Koh, I. Y. Y., Alexov, E., and Honig, B. (2003). Using multiple structure alignments, fast model building, and energetic analysis in fold recognition and homology modeling. *Proteins*, 53 Suppl 6:430–435.

Phillips, D. C. (1970). The development of crystallographic enzymology. *Biochem Soc Symp*, 30:11–28.

Pleasance, E. D., Cheetham, R. K., Stephens, P. J., McBride, D. J., Humphray, S. J., Greenman, C. D., Varela, I., Lin, M. L., Ordonez, G. R., Bignell, G. R., Ye, K., Alipaz, J., Bauer, M. J., Beare, D., Butler, A., Carter, R. J., Chen, L., Cox, A. J., Edkins, S., Kokko-Gonzales, P. I., Gormley, N. A., Grocock, R. J., Haudenschild, C. D., Hims, M. M., James, T., Jia, M., Kingsbury, Z., Leroy, C., Marshall, J., Menzies, A., Mudie, L. J., Ning, Z., Royce, T., Schulz-Trieglaff, O. B., Spiridou, A., Stebbings, L. A., Szajkowski, L., Teague, J., Williamson, D., Chin, L., Ross, M. T., Campbell, P. J., Bentley, D. R., Futreal, P. A., and Stratton, M. R. (2010a). A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature*, 463(7278):191–6.

Pleasance, E. D., Stephens, P. J., O'Meara, S., McBride, D. J., Meynert, A., Jones, D., Lin, M. L., Beare, D., Lau, K. W., Greenman, C., Varela, I., Nik-Zainal, S., Davies, H. R., Ordonez, G. R., Mudie, L. J., Latimer, C., Edkins, S., Stebbings, L., Chen, L., Jia, M., Leroy, C., Marshall, J., Menzies, A., Butler, A., Teague, J. W., Mangion, J., Sun, Y. A., McLaughlin, S. F., Peckham, H. E., Tsung, E. F., Costa, G. L., Lee, C. C., Minna, J. D., Gazdar, A., Birney, E., Rhodes, M. D., McKernan, K. J., Stratton, M. R., Futreal, P. A., and Campbell, P. J. (2010b). A small-cell lung cancer genome with complex signatures of tobacco exposure. *Nature*, 463(7278):184–90.

Pollastri, G. and Baldi, P. (2002). Prediction of contact maps by giohmms and recurrent neural networks using lateral propagation from all four cardinal corners. *Bioinformatics*, 18 Suppl 1:S62–S70.

Pollock, D. D., Taylor, W. R., and Goldman, N. (1999). Coevolving protein residues: maximum likelihood identification and relationship to structure. *J Mol Biol*, 287(1):187–198.

Ponder, J. W. and Richards, F. M. (1987a). An efficient newton-like method for molecular mechanics energy minimization of large molecules. *Journal of Computational Chemistry*, 8:1016–1024. Tinker main reference.

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

Porter, C. T., Bartlett, G. J., and Thornton, J. M. (2004). The catalytic site atlas: a resource of catalytic sites and residues identified in enzymes using structural data. *Nucleic Acids Res*, 32(Database issue):D129–33.

Potapov, V., Cohen, M., and Schreiber, G. (2009). Assessing computational methods for predicting protein stability upon mutation: good on average but not in the details. *Protein Eng Des Sel*, 22(9):553–60.

Punta, M. and Rost, B. (2005a). PROFcon: novel prediction of long-range contacts. *Bioinformatics*, 21(13):2960–2968.

Punta, M. and Rost, B. (2005b). Protein folding rates estimated from contact predictions. *J Mol Biol*, 348(3):507–512.

Qian, B., Raman, S., Das, R., Bradley, P., McCoy, A. J., Read, R. J., and Baker, D. (2007). High-resolution structure prediction and the crystallographic phase problem. *Nature*, 450(7167):259–264.

Raman, S., Huang, Y. J., Mao, B., Rossi, P., Aramini, J. M., Liu, G., Montelione, G. T., and Baker, D. (2010a). Accurate automated protein nmr structure determination using unassigned noesy data. *J Am Chem Soc*, 132(1):202–207.

Raman, S., Lange, O. F., Rossi, P., Tyka, M., Wang, X., Aramini, J., Liu, G., Ramelot, T. A., Eletsky, A., Szyperski, T., Kennedy, M. A., Prestegard, J., Montelione, G. T., and Baker, D. (2010b). Nmr structure determination for larger proteins using backbone-only data. *Science*, 327(5968):1014–1018.

Raman, S., Qian, B., Baker, D., and Walker, R. C. (2008). Advances in rosetta protein structure prediction on massively parallel systems. *IBM Journal of Research and Development*, 52(1):7–17.

Ramensky, V., Bork, P., and Sunyaev, S. (2002). Human non-synonymous snps: server and survey. *Nucleic acids research*, 30(12202775):3894–900.

Ripka, W. C. (1986). Computer-assisted model building. *Nature*, 321:93–94.

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

Röthlisberger, D., Khersonsky, O., Wollacott, A. M., Jiang, L., DeChancie, J., Betker, J., Gallaher, J. L., Althoff, E. A., Zanghellini, A., Dym, O., Albeck, S., Houk, K. N., Tawfik, D. S., and Baker, D. (2008). Kemp elimination catalysts by computational enzyme design. *Nature*, 453(7192):190–195.

Russell, R. B. and Barton, G. J. (1992). Multiple protein sequence alignment from tertiary structure comparison: assignment of global and residue confidence levels. *Proteins*, 14(2):309–323.

Russo, E. and Bunk, S. (1999). Hot papers in bioinformatics. *The Scientist*, (8).

Sali, A. and Blundell, T. L. (1990). Definition of general topological equivalence in protein structures. a procedure involving comparison of properties and relationships through simulated annealing and dynamic programming. *J Mol Biol*, 212(2):403–428.

Sali, A. and Blundell, T. L. (1993). Comparative protein modelling by satisfaction of spatial restraints. *J Mol Biol*, 234(3):779–815.

Sali, A., Shakhnovich, E., and Karplus, M. (1994). How does a protein fold? *Nature*, 369(6477):248–251.

Sathyapriya, R., Duarte, J. M., Stehr, H., Filippis, I., and Lappe, M. (2009). Defining an essence of structure determining residue contacts in proteins. *PLoS Comput Biol*, 5(12):e1000584.

Schenk, G., Margraf, T., and Torda, A. E. (2008). Protein sequence and structure alignments within one framework. *Algorithms Mol Biol*, 3:4.

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

Schneidman-Duhovny, D., Hammel, M., and Sali, A. (2011). Macromolecular docking restrained by a small angle x-ray scattering profile. *J Struct Biol*, 173(3):461–471.

Schwede, T., Kopp, J., Guex, N., and Peitsch, M. C. (2003). Swiss-model: An automated protein homology-modeling server. *Nucleic Acids Res*, 31(13):3381–3385.

Scouloudi, H. (1959). The myoglobin molecule. *Nature*, 183(4658):374–376.

Shackelford, G. and Karplus, K. (2007). Contact prediction using mutual information and neural nets. *Proteins*, 69 Suppl 8:159–164.

Shao, Y. and Bystroff, C. (2003). Predicting interresidue contacts using templates and pathways. *Proteins*, 53 Suppl 6:497–502.

Shatsky, M., Nussinov, R., and Wolfson, H. J. (2004). A method for simultaneous alignment of multiple protein structures. *Proteins*, 56(1):143–156.

Shen, Y., Lange, O., Delaglio, F., Rossi, P., Aramini, J. M., Liu, G., Eletsky, A., Wu, Y., Singarapu, K. K., Lemak, A., Ignatchenko, A., Arrowsmith, C. H., Szyperski, T., Montelione, G. T., Baker, D., and Bax, A. (2008). Consistent blind protein structure generation from nmr chemical shift data. *Proc Natl Acad Sci U S A*, 105(12):4685–4690.

Sherry, S. T., Ward, M. H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., and Sirotkin, K. (2001). dbsnp: the ncbi database of genetic variation. *Nucleic Acids Res*, 29(1):308–11.

Shindyalov, I. N. and Bourne, P. E. (1998). Protein structure alignment by incremental combinatorial extension (ce) of the optimal path. *Protein Eng*, 11(9):739–747.

Shrake, A. and Rupley, J. A. (1973). Environment and exposure to solvent of protein atoms. lysozyme and insulin. *J Mol Biol*, 79(2):351–371.

Shu, H. K., Pelley, R. J., and Kung, H. J. (1990). Tissue-specific transformation by epidermal growth factor receptor: a single point mutation within the atp-binding pocket of the erbb product increases its intrinsic kinase activity and activates its sarcomagenic potential. *Proc Natl Acad Sci U S A*, 87(23):9103–7.

Simons, K. T., Kooperberg, C., Huang, E., and Baker, D. (1997). Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and bayesian scoring functions. *J Mol Biol*, 268(1):209–225.

Singer, M. S., Vriend, G., and Bywater, R. P. (2002). Prediction of protein residue contacts with a pdb-derived likelihood matrix. *Protein Eng*, 15(9):721–725.

Sippl, M. J. (1995). Knowledge-based potentials for proteins. *Curr Opin Struct Biol*, 5(2):229–235.

Sjoblom, T., Jones, S., Wood, L. D., Parsons, D. W., Lin, J., Barber, T. D., Mandelker, D., Leary, R. J., Ptak, J., Silliman, N., Szabo, S., Buckhaults, P., Farrell, C., Meeh, P., Markowitz, S. D., Willis, J., Dawson, D., Willson, J. K., Gazdar, A. F., Hartigan, J., Wu, L., Liu, C., Parmigiani, G., Park, B. H., Bachman, K. E., Papadopoulos, N., Vogelstein, B., Kinzler, K. W., and Velculescu, V. E. (2006). The consensus coding sequences of human breast and colorectal cancers. *Science*, 314(5797):268–74.

Smith, T. F. and Waterman, M. S. (1981). Identification of common molecular subsequences. *J Mol Biol*, 147(1):195–197.

Snow, C. D., Nguyen, H., Pande, V. S., and Gruebele, M. (2002). Absolute comparison of simulated and experimental protein-folding dynamics. *Nature*, 420(6911):102–106.

Soto, C. S., Fasnacht, M., Zhu, J., Forrest, L., and Honig, B. (2008). Loop modeling: Sampling, filtering, and scoring. *Proteins*, 70(3):834–843.

Sugita, Y. and Okamoto, Y. (1999). Replica-exchange molecular dynamics method for protein folding. *Chemical Physics Letters*, 314:141–151.

Sutcliffe, M. J., Haneef, I., Carney, D., and Blundell, T. L. (1987). Knowledge based modelling of homologous proteins, part i: Three-dimensional frameworks derived from the simultaneous superposition of multiple structures. *Protein Eng*, 1(5):377–384.

Talavera, D., Taylor, M. S., and Thornton, J. M. (2010). The (non)malignancy of cancerous amino acidic substitutions. *Proteins*, 78(3):518–29.

Tanaka, S. and Scheraga, H. A. (1976). Medium- and long-range interaction parameters between amino acids for predicting three-dimensional structures of proteins. *Macromolecules*, 9(6):945–950.

Taylor, T. J. and Vaisman, I. I. (2006). Graph theoretic properties of networks formed by the delaunay tessellation of protein structures. *Phys. Rev. E*, 73(4):041925.

Tegge, A. N., Wang, Z., Eickholt, J., and Cheng, J. (2009). NNcon: improved protein contact map prediction using 2D-recursive neural networks. *Nucleic Acids Res*, 37(Web Server issue):W515–W518.

The UniProt Consortium (2010). The Universal Protein Resource (UniProt) in 2010. *Nucl. Acids Res.*, 38(suppl1):D142–148.

Torda, A. E., Procter, J. B., and Huber, T. (2004). Wurst: a protein threading server with a structural scoring function, sequence profiles and optimized substitution matrices. *Nucleic Acids Res*, 32(Web Server issue):W532–W535.

Torkamani, A. and Schork, N. J. (2008). Prediction of cancer driver mutations in protein kinases. *Cancer Res*, 68(6):1675–82.

Tress, M. L. and Valencia, A. (2010). Predicted residue-residue contacts can help the scoring of 3D models. *Proteins*, 78(8):1980–1991.

Tyagi, M., Gowri, V. S., Srinivasan, N., de Brevern, A. G., and Offmann, B. (2006). A substitution matrix for structural alphabet based on structural alignment of homologous proteins and its applications. *Proteins*, 65(1):32–39.

Unger, R., Harel, D., Wherland, S., and Sussman, J. L. (1989). A 3d building blocks approach to analyzing and predicting structure of proteins. *Proteins*, 5(4):355–373.

Vendruscolo, M., Kussell, E., and Domany, E. (1997). Recovery of protein structure from contact maps. *Fold Des*, 2(5):295–306.

Veretnik, S., Bourne, P. E., Alexandrov, N. N., and Shindyalov, I. N. (2004). Toward consistent assignment of structural domains in proteins. *J Mol Biol*, 339(3):647–678.

Vicatos, S., Reddy, B. V. B., and Kaznessis, Y. (2005). Prediction of distant residue contacts with the use of evolutionary information. *Proteins*, 58(4):935–949.

Vincent, J. J., Tai, C.-H., Sathyanarayana, B. K., and Lee, B. (2005). Assessment of Casp6 predictions for new and nearly new fold targets. *Proteins*, 61 Suppl 7:67–83.

Vogelstein, B. and Kinzler, K. W. (1993). The multistep nature of cancer. *Trends Genet*, 9(4):138–141.

Vullo, A., Walsh, I., and Pollastri, G. (2006). A two-stage approach for improved prediction of residue contact maps. *BMC Bioinformatics*, 7:180.

Wang, Z. and Moult, J. (2001). Snps, protein structure, and disease. *Human Mutation*, 17(11295823):263–70.

Wang, Z. X. (1996). How many fold types of protein are there in nature? *Proteins*, 26(2):186–191.

Wang, Z. X. (1998). A re-estimation for the total numbers of protein folds and superfamilies. *Protein Eng*, 11(8):621–626.

Westhead, D. R., Collura, V. P., Eldridge, M. D., Firth, M. A., Li, J., and Murray, C. W. (1995). Protein fold recognition by threading: comparison of algorithms and analysis of results. *Protein Eng*, 8(12):1197–1204.

Wheeler, D. L., Barrett, T., Benson, D. A., Bryant, S. H., Canese, K., Chetvernin, V., Church, D. M., Dicuccio, M., Edgar, R., Federhen, S., Feolo, M., Geer, L. Y., Helmberg, W., Kapustin, Y., Khovayko, O., Landsman, D., Lipman, D. J., Madden, T. L., Maglott, D. R., Miller, V., Ostell, J., Pruitt, K. D., Schuler, G. D., Shumway,

M., Sequeira, E., Sherry, S. T., Sirotkin, K., Souvorov, A., Starchenko, G., Tatusov, R. L., Tatusova, T. A., Wagner, L., and Yaschenko, E. (2008). Database resources of the national center for biotechnology information. *Nucleic Acids Res*, 36(Database issue):D13–21.

Winker, S., Overbeek, R., Woese, C. R., Olsen, G. J., and Pfluger, N. (1990). Structure detection through automated covariance search. *Comput Appl Biosci*, 6(4):365–371.

Wohlers, I., Petzold, L., Domingues, F., and Klau, G. (2009). Paul: protein structural alignment using integer linear programming and lagrangian relaxation. *BMC Bioinformatics*, 10(Suppl 13):P2.

Wood, L. D., Parsons, D. W., Jones, S., Lin, J., Sjoblom, T., Leary, R. J., Shen, D., Boca, S. M., Barber, T., Ptak, J., Silliman, N., Szabo, S., Dezso, Z., Ustyanksky, V., Nikolskaya, T., Nikolsky, Y., Karchin, R., Wilson, P. A., Kaminker, J. S., Zhang, Z., Croshaw, R., Willis, J., Dawson, D., Shipitsin, M., Willson, J. K., Sukumar, S., Polyak, K., Park, B. H., Pethiyagoda, C. L., Pant, P. V., Ballinger, D. G., Sparks, A. B., Hartigan, J., Smith, D. R., Suh, E., Papadopoulos, N., Buckhaults, P., Markowitz, S. D., Parmigiani, G., Kinzler, K. W., Velculescu, V. E., and Vogelstein, B. (2007). The genomic landscapes of human breast and colorectal cancers. *Science*, 318(5853):1108–13.

Wu, H. (1995). Studies on denaturation of proteins. xiii. a theory of denaturation. 1931. *Adv Protein Chem*, 46:6–26; discussion 1–5.

Wu, S., Skolnick, J., and Zhang, Y. (2007). Ab initio modeling of small proteins by iterative tasser simulations. *BMC Biol*, 5:17.

Xie, W. and Sahinidis, N. V. (2007). A reduction-based exact algorithm for the contact map overlap problem. *J Comput Biol*, 14(5):637–654.

Xu, J., Jiao, F., and Berger, B. (2005). A tree-decomposition approach to protein structure prediction. In *Proc. IEEE Computational Systems Bioinformatics Conf*, pages 247–256.

Xu, J., Li, M., Kim, D., and Xu, Y. (2003). Raptor: optimal protein threading by linear programming. *J Bioinform Comput Biol*, 1(1):95–117.

Xu, Y. and Xu, D. (2000). Protein threading using prospect: design and evaluation. *Proteins*, 40(3):343–354.

Xu, Y., Xu, D., Gabow, H. N., and Gabow, H. (2000). Protein domain decomposition using a graph-theoretic approach. *Bioinformatics*, 16(12):1091–1104.

Xue, B., Faraggi, E., and Zhou, Y. (2009). Predicting residue-residue contact maps by a two-layer, integrated neural-network method. *Proteins*, 76(1):176–183.

Ye, J. and Janardan, R. (2004). Approximate multiple protein structure alignment using the sum-of-pairs distance. *J Comput Biol*, 11(5):986–1000.

Ye, J., Pavlicek, A., Lunney, E. A., Rejto, P. A., and Teng, C. H. (2010). Statistical method on nonrandom clustering with application to somatic mutations in cancer. *BMC Bioinformatics*, 11:11.

Yue, P., Forrest, W. F., Kaminker, J. S., Lohr, S., Zhang, Z., and Cavet, G. (2010). Inferring the functional effects of mutation through clusters of mutations in homologous proteins. *Hum Mutat*, 31(3):264–71.

Zemla, A. (2003). LGA: A method for finding 3D similarities in protein structures. *Nucleic Acids Res*, 31(13):3370–3374.

Zhang, C. and DeLisi, C. (1998). Estimating the number of protein folds. *J Mol Biol*, 284(5):1301–1305.

Zhang, Y. and Skolnick, J. (2004). Automated structure prediction of weakly homologous proteins on a genomic scale. *Proc Natl Acad Sci U S A*, 101(20):7594–7599.

Zhang, Y. and Skolnick, J. (2005). The protein structure prediction problem could be solved using the current pdb library. *Proc Natl Acad Sci U S A*, 102(4):1029–1034.

Zhao, Y. and Karypis, G. (2003). Prediction of contact maps using support vector machines. In *Proc. Third IEEE Symp. Bioinformatics and Bioengineering*, pages 26–33.

Zhou, H. and Zhou, Y. (2002). Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. *Protein Sci*, 11(11):2714–2726.

Zwanzig, R., Szabo, A., and Bagchi, B. (1992). Levinthal's paradox. *Proc Natl Acad Sci U S A*, 89(1):20–22.