Strategies for the structure-based analysis of protein functionality

Dissertation zur Erlangung des akademischen Grades des

Doktors der Naturwissenschaften (Dr. rer. nat.)

eingereicht im Fachbereich Biologie, Chemie, Pharmazie

der Freien Universität Berlin

vorgelegt in englischer Sprache von

Aysam Gürler

aus Bonn

Juni, 2009

Diese Dissertation wurde von der Deutschen Forschungsgemeinschaft finanziert und am Institut für Chemie und Biochemie der Freien Universität Berlin in englischer Sprache verfasst.

1. Gutachter: Prof. Ernst-Walter Knapp
2. Gutachter: Prof. Wolfram Saenger

Disputation am 24.06.2009

Preamble

This cumulative thesis is the sum of my research work, regarding the structural analysis of protein functionality. The focus lies on the similarity of protein structures to each other. This thesis is based on the following three peer-reviewed journal publications:

Guerler A, Knapp EW
Novel folds and their non-sequential structural analogs
Protein Science 17:8, 1374-82, 2008

Bauer R, Bourne PE, Formella A, Frömmel C, Gille C, Goede A, Guerler A, Hoppe A, Knapp, EW, Pöschel T, Wittig B, Ziegler V, Preissner R
Superimposé: A 3D structural superposition server
Nucleic Acids Research 36, W47-W54, 2008

Guerler A, Knapp EW
Evaluation of sequence alignments of distantly related sequence pairs with respect to structural similarity Genome Informatics 18, 183-91, 2007

During my PhD research, additionally the following two papers were published, which illustrate the research results on molecular interference of protein structures with other proteins, respectively small molecules:

Guerler A, Lorenzen S, Krull F, Knapp EW
Sampling geometries of protein-protein complexes
Genome Informatics 20, 260-9, 2008

Guerler A*, Moll S, Weber M, Meyer H, Cordes F
Selection and flexible optimization of binding modes from conformation ensembles
Elsevier BioSystems 92, 42-8, 2008

Acknowledgements

I would like to thank Prof. Ernst-Walter Knapp, Prof. Wolfram Saenger and my colleagues and friends Christoph Gille, Raphael Bauer, Jorge Numata, Jonas Maaskola and many others for their valuable support. The following work was funded by the International Research Training Group (IRTG) on “Genomics and Systems Biology of Molecular Networks” (GRK1360, Deutsche Forschungsgemeinschaft (DFG)).

Contents

1 INTRODUCTION

Most cellular processes of life are regulated and accomplished by proteins. The production of proteins itself is such a process, which is initiated in the nucleus. The nucleus encapsulates the DNA (deoxyribonucleic acid), which is a linear chain molecule consisting of a sequence of four different bases, namely Adenine (A), Cytosine (C), Guanine (G) and Thymine (T). Their sequence encodes a cell's genes (genome) as a four letter code. The genes themselves can be regarded as instructions for the construction of proteins or other molecules, which are post-processed or do directly fulfill a certain regulatory function. Figure 1.1 shows a schematic illustration of the major events in the generation of biologically active proteins. Initially, a DNA segment (gene) encoding a polypeptide chain is transcribed and spliced to mRNA (messenger ribonucleic acid). The mRNA is translated by the ribosomes at the endoplasmatic reticulum into a polypeptide chain. Due to energetic preferences i.e. hydrophobicity the polypeptide chain folds to a biologically active protein. In some cases, additional post-processing might be required

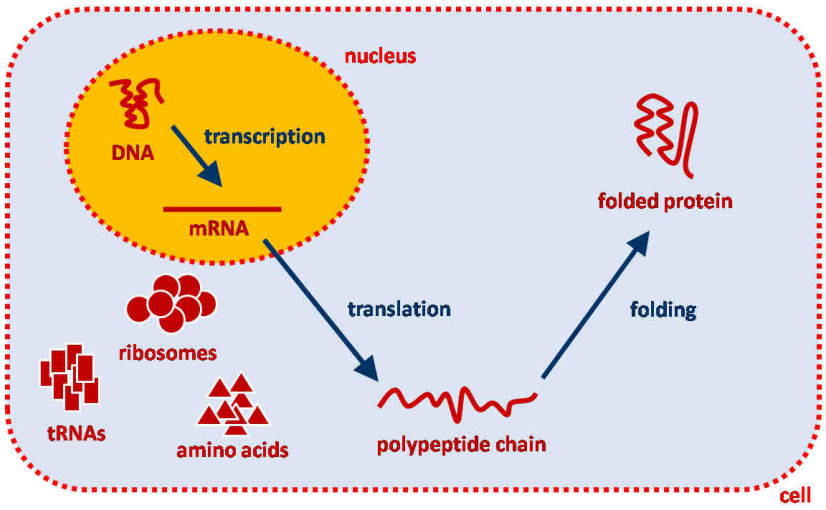


Figure 1.1 Illustration of the major processes and ingredients in the generation of protein folds. A gene of the DNA is transcribed to mRNA (messenger RNA), which is translated to a polypeptide chain by the ribosomes with the usage of tRNAs and amino acids. Finally, the polypeptide chain folds to the active protein structure, due to energetic preferences.

to activate the protein i.e. placement of cofactors, improvement and alteration of fold i.e. by Chaperone (Buchner 2002). The genetic regions, which encode proteins are called “coding regions”. In this context, “non-coding” DNA sequences may refer to sites with regulatory functions for the expression of “coding” genes. Despite this, recent research clearly illustrated that certain fractions of the “non-coding” sequences encode not proteins, but µRNAs (He and Hannon 2004), which directly fulfill very important regulatory functions. In case of proteins each triplet of DNA bases, encodes one out of 20 amino acids. These are organic compounds. To form a protein, the amino acids are arranged in a linear chain and joined together by peptide bonds between the carboxyl and amino groups of adjacent amino acid residues. All amino acids share structural properties, containing an α-carbon, an amino group, a carboxyl group and a variable side chain. Only proline is different, due to a ring formed with the N-end amine group. The side chains have different physico-chemical properties, which affect the 3D dimensional protein structure and thereby the functionality of certain protein sites. Each protein has a final negatively charged carboxyl-group at its C-terminus and a free positively charged amino-group at its N-terminus.

Proteins are involved in cell signaling, immune response, cell adhesion, proliferation and many other processes. There are numerous proteins, which catalyze biochemical processes and are vital to the metabolism in a living cell. Also structural or mechanical functions are fulfilled with proteins i.e. actin and myosin in muscles or other proteins stabilizing the cytoskeleton. In 1838, the Swedish chemist Jöns Jakob Berzelius appeared to be the first to describe proteins (Hartley 1951). The term protein is derived from the Greek word “proteios”, which means “primary”. However, their importance and impact has not been noticed until 1926 when it was found that the urease is a protein (JB 1926). In 1958, Perutz (Muirhead and Perutz 1963) and Kendrew (Kendrew et al. 1958) were the first to apply the x-ray crystallography to resolve the atomic details of whole proteins. Today, the protein database (PDB) (Berman et al. 2000) contains the heavy atom coordinates of about 50.000 protein structures.

The specific biochemical abilities of a protein result from its three-dimensional (3D) native structure. Hereby, four different structural levels are distinguished. The primary structure is defined by the amino acid sequence. The secondary structure level is defined by structural units i.e. α-helix, β-sheet, 3^{10}-helix and others, stabilized by hydrogen bonds between the protein backbone atoms. The third level is the tertiary structure, which defines the complete fold of a single polypeptide chain. Proteins can also work together to achieve a particular function, and they often associate to form stable complexes (quaternary structure). For enzymes the structure optimizes the geometric arrangement of catalytically active amino acid side chains and cofactors and simultaneously allows efficient access and removal of educts and products. For proteins where one of the functions is to form specific complexes with other proteins, the shape of the contact surface and the residue pair interactions in the contact surface are also relevant (Shulman-Peleg et al. 2007). Consequently, proteins with the same function often have the same structure and key residues involved in the various aspects of function are conserved among different species. However, there are exceptions where nature uses alternatively designed protein 3D structures with equivalent or different key residues and cofactors to perform the same function in different species.

Under physiological conditions the native 3D structure of a protein is determined solely by the primary sequence (Anfinsen et al. 1961). On the other hand the native 3D structure does not belong to a unique primary sequence. Mutational studies demonstrated that often only a small fraction of amino acids is crucial to define and stabilize the 3D structures of proteins (Guo et al. 2004; Russ et al. 2006). Consequently, only structurally and functionally relevant residues of a protein are conserved among different species. This fact is used to assess the unknown function of proteins by sequence comparisons, which may fail if the sequence homology is too low. In case the protein 3D structure is available, structure comparison can be more useful to assess the protein function, since the universe of protein structures is much smaller than the universe of protein sequences. The number of different protein folds is estimated

to be about only 1000 (Wang 1998), but there are also less optimistic views on the number of distinct protein folds (Grant et al. 2004).

Polypeptide sequence comparison is now routine work, if sequence similarity is sufficiently high as for instance more than 40% identity. However, it becomes increasingly uncertain with lower sequence identity (Guerler and Knapp 2007). Since not only sequence but also structure similarity of proteins correlates with their function, structure comparison of proteins is most useful to characterize a protein of yet unknown function, if its 3D structure is available. This approach can be particularly successful, since at present the Protein Data Bank (PDB) (Berman, 2000) contains already a considerable fraction of the universe of folds to predict the structures of soluble proteins (Kolodny et al. 2005).

Protein 3D structure comparison is still a challenging task and depends heavily on the alignment algorithm, the similarity measure used and on the fractions of the protein structures considered for the pairwise structure alignment (Kolodny, 2005). An actual but still incomplete listing of available methods for protein structure alignment can be found on the webpage *http://en.wikipedia.org/wiki/Structural_alignment_software* containing more than 40 different programs.

To identify a structural similarity between protein structures, the considered protein structure must be aligned to all representative protein structures of the PDB. Suitable databases of representative protein structures are the ASTRAL databases provided by SCOP (Chandonia et al. 2004). These databases contain subsets of the PDB with domain structures whose sequence similarity is below a threshold value of say 40% or 70% sequence identity. In the past the structure alignment methods used to identify the same fold in a database were often restricted or biased by only considering protein structures, which possess the same connectivity of secondary structure elements (SSEs) (i.e. α-helices and β-strands), as defined by the polypeptide chain.

Only few methods are available that allow for non-sequential protein structure alignments that can be combined with a database of protein structures (Fischer, 1996; Yuan, 2005; Shih, 2006). But, most of these methods offer only to scan a database of pre-calculated protein structure alignments and do not allow aligning a yet unknown protein structure to a database of protein structures. Recently the program GANGSTA (Kolbeck et al. 2006) appeared, which ignores the loops connecting different SSEs and therefore allows unbiased non-sequential protein structure alignment. In addition, it offers alignment of yet unknown protein structures against a database of more than 3,000 domains of protein structures with less than 40% sequence identity. Since GANGSTA is relatively slow, we have redesigned it completely. GANGSTA+ is more than a factor of ten faster, yields alignments of higher quality and offers alignments of arbitrary protein structures against the ASTRAL40 (Chandonia et al. 2004) (1.71) database containing more than 7,000 domains of protein structures with less than 40% sequence identity to each other. Similar to the former GANGSTA algorithm, GANGSTA+ initially maximizes the contact map overlap of an SSE alignment (first phase), before refining the result on residue-level (second phase). However, the algorithms to achieve a high contact map overlap and a fast refinement on residue level are entirely different. In contrast to a stochastically genetic algorithm and a greedy search, GANGSTA+ uses a deterministic combinatorial approach (first phase), and a physical point matching approach (Guerler et al. 2008b) for the refinement on residue-level. The point matching approach optionally allows the detection of SSEs aligned in reverse sequence direction. GANGSTA+ has two essential parameters regarding the search depth of the two phases (see supplement of (Guerler and Knapp 2008) for details). It is found, that the refinement phase (second phase) is very important for the robustness of GANGSTA+. Hence, we took advantage of the fast point matching approach and increased the number of refined SSE alignments from 10 to 200 with regard to GANGSTA. GANGSTA+ still remained fast (~1s per structure alignment on an ATHLON 1.6 GHz). Beyond that GANGSTA+ has a third phase where the align-

ment is enlarged to coiled regions. We analyzed about 50 million protein structure pairs with GANGSTA+ and made the results available online (*http://agknapp.chemie.fu-berlin.de/gplus*).

2 PUBLICATIONS

2.1 Novel protein folds and their non-sequential structural analogs

Authors Guerler A, Knapp EW

Bibliography Protein Science 17:8, 1374-82, 2008

Contribution

- Development of the research question
- Development of the required software
- Generation and analysis of the results
- Manuscript preparation

http://dx.doi.org/10.1110/ps.035469.108

In this publication GANGSTA+ is described in detail and has been applied to align newly determined protein folds with no known homologues and the ASTRAL40 (SCOP version 1.71) database (Murzin et al. 1995). It is demonstrated that protein folds, which are considered to be new, appear to be known folds, if one considers only the topological arrangement of the SSEs and disregards the connectivity of the polypeptide chain defined by the loops connecting the SSEs. GANGSTA+ is capable to perform also structure alignments where the SSEs can be aligned in reverse orientation, i.e. aligned SSE pairs are of the same type, but oriented such that the C-terminal end of one SSE is superimposed on the N-terminal end of the other SSE. This can be used to enhance the likeability to find similar structures for a given protein structure. To contrast non-sequential with sequential alignment results, we applied DaliLite. Figure 2.1 illustrates the results for 2ES9 (Benach et al. 2005). A search with DaliLite for structures similar to 2ES9, yielded a sequential structure alignment of significant similarity (Z-score > 2.0) to the structure of 1SZA (Meinhart and Cramer 2004). DaliLite aligns 67 residues at RMSD = 2.5 Å for this protein pair. However, the generated alignment is insufficient to describe the fold of 2ES9 in sufficient detail, since only the four-helix bundle, which is a common motive (Mehl et al. 2003), was aligned, while the fifth lateral α-helix was skipped. The structure alignment with GANGSTA+ for 2ES9 with respect to the ASTRAL40 database took 40 min (~0.3 s per protein pair) and revealed a non-sequential alignment with 1SXJ, involving all five SSEs of 2ES9 (see Fig. 2.1) with 69 aligned residues at 1.8 Å RMSD. Given the protein pair 2ES9 and 1SXJ, DaliLite aligned 57 residues at 2.5 Å RMSD with a Z-score of 3.2 > 2.0. It succeeded in aligning the four-helix bundle, but again skipped the lateral α-helix (see Fig. 2.1). This demonstrates that GANGSTA+ is able to detect non-sequential similarities for protein chains stated to possess new folds. Although the question whether a new protein structure contains a new fold or not remains difficult to judge, the results illustrate that the application of non-sequential structure alignment tools can yield additional insight to understand protein structures and fold characteristics, presenting new starting points for protein function analysis and protein structure comparison. GANGSTA+ is not

bound to a sequential connectivity of SSEs in the polypeptide chains of proteins. Thus, it can detect structure similarities of different proteins that have common ancestors, but whose SSE connectivity was reshuffled by genetic operations (Cooper et al. 1997).

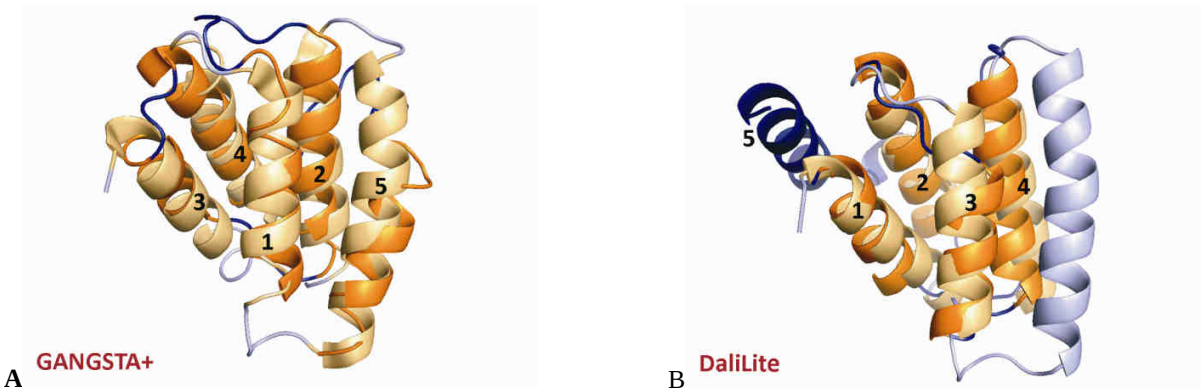


Figure 2.1. (A) Protein structure alignment with GANGSTA+. New fold **2ES9** (Benach et al. 2005) aligned on **1SXJ** (Bowman et al. 2004) yielding the RMSD = 1.8 Å with 69 aligned residues and 5 aligned SSEs. The aligned SSEs of **2ES9** (**1SXJ**) are represented in dark (light) orange, not aligned parts (SSEs and loops) are in dark (light) blue. **(B)** Protein structure alignment with DaliLite. New fold **2ES9** (Benach et al. 2005) aligned on **1SXJ** (Bowman et al. 2004) yielding the RMSD = 2.5 Å with 57 aligned residues and 4 aligned SSEs. The aligned SSEs of **2ES9** (**1SXJ**) are represented in dark (light) orange, not aligned parts (SSEs and loops) are in dark (light) blue. The fifth lateral α-helix of **2ES9** has not been aligned.

2.2 Superimposé: A 3D structural superposition server

Authors Bauer R, Bourne PE, Formella A, Frömmel C, Gille C, Goede A, Guerler A, Hoppe A, Knapp, EW, Pöschel T, Wittig B, Ziegler V, Preissner R

Bibliography Nucleic Acids Research 36, W47-W54, 2008

Contribution

- Supply of the non-sequential structure alignment software
- Visualization of the results of the provided software
- Technical support in application and server embedding
- Generation and analysis of results
- Contribution to manuscript preparation

http://dx.doi.org/10.1093/nar/gkn285

In this publication, the 3D structure alignment server Superimposé is described. The server comprises several algorithms for the structural alignment of small and large molecules against several commonly used databases as for instance ASTRAL (Murzin et al. 1995). We provided GangstaLite (Bauer et al. 2008), which is the only non-sequential structure alignment method of the server. For Superimposé, it has been decided to provide a wizard style approach that guides the user through different possibilities offered. For all algorithms a fixed set of parameters is defined, allowing a generalized execution task. A typical search workflow begins with the selection of a task the user wants to execute. Figure 2.2 illustrates the online visualization of the results after aligning a protein structure against the ASTRAL40 (SCOP version 1.69) database.

The Superimposé server will be useful for bioinformaticians who have specialized on structures, macromolecular biologists and the systems biology community by providing possibilities to identify similar patches (binding sites/surface patches) in known proteins. By reducing the complexity of installing algorithms, databases and defining suitable parameter sets Superimposé allows researchers to instantly deal with the task without the administrative problems around it. In the near future, additional feature and appropriate server extensions will be made available.

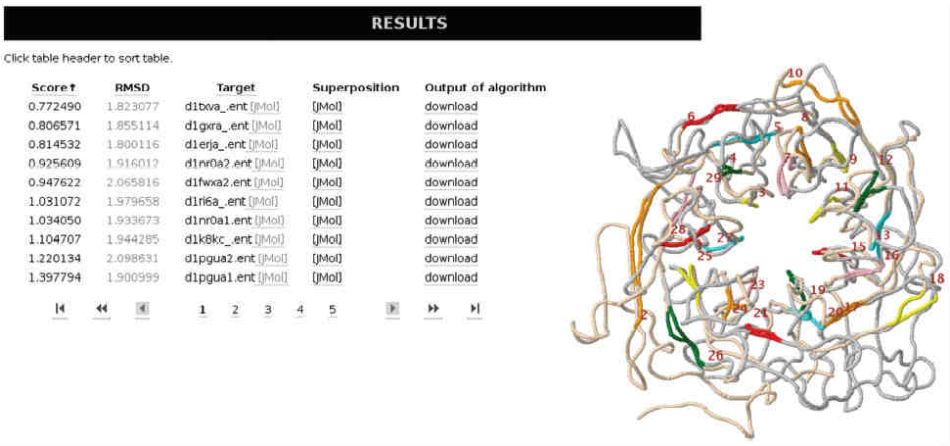


Figure 2.2 Superimposé results (left) and non-sequential structural alignment generated by GangstaLite (right).

2.3 Evaluation of sequence alignments of distantly related sequence pairs with respect to structural similarity

Authors Guerler A, Knapp EW

Bibliography Genome Informatics 18, 183-91, 2007

Contribution

- Development of the research question
- Development of the required software
- Generation and analysis of the results
- Manuscript preparation

The full article is attached to the end of this document.

In this publication, the performance of common substitution matrices of sequence alignment methods (Henikoff and Henikoff 1992) in detecting structural similarities is evaluated with the ASTRAL40 (SCOP version 1.69) database (Murzin et al. 1995). The database consists of 7290 protein chains, which share less than 40% sequence identity. Initially, a structure alignment database (SD) has been set up by evaluation of all ASTRAL40 pairs, which leads to about 26 million structural alignments. Only the highest scoring alignment of each pair with a structural score ($SC = 100 - 200 * RMSD / N_{aligned}$) above 30 and at least 50% of the secondary structure elements in the smaller of both proteins aligned are kept. This amounts to about 18.6 million protein pairs. From them, about 450.000 pairs have a structural score above 90 SC. Thus on average, each ASTRAL40 entry shares very high structural similarities with about 60 other proteins. About 7.15 million pairs score above 80 SC, which indicates significant structural similarities between each ASTRAL40 entry and 980 other proteins on average (about 13% of the ASTRAL40 set).

Then, the sequences of each ASTRAL40 entry is aligned on the complete sequence set with FASTA. The list of the 100 highest ranked protein pairs for each entry (as SCOP 1.69 codes) is used to determine the corresponding structural scores of the structure alignment database. This procedure is applied in combination with BLOSUM50, BLOSUM62 and PAM120. Additionally, the 100 highest structural scores for each ASTRAL40 entry are selected from our structure alignment database and plotted as

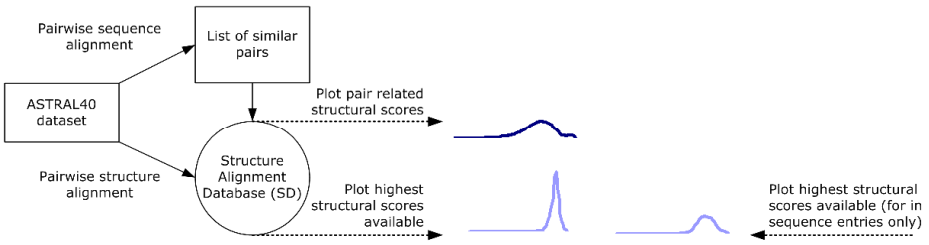


Figure 2.3. This figure illustrates the data acquisition process by usage of sequence (dark) and structure (light) alignments. As result the structural score distributions, according to the structural alignment database (SD), are plotted. Additionally, the sequential structure alignment entries are plotted separately.

reference or as upper performance limit. Since our structure alignment method is able to detect non-sequential similarities between two protein structures, we additionally plotted the sequential structure alignments separately (see figure 2.3). The resulting structural scores are plotted in figure 2.4 and illustrate the difficulties of sequence alignment approaches in cases of low sequence similarity to already known protein structures. The sequence alignment method is able to reproduce the structurally most similar protein pairs, but in 25% of all high ranking FASTA results only very little structural similarity could be detected. This is related to the simplification of the model, since the sequence alignment method only incorporates the primary structure. Additionally, the sequence alignment method employs substitution matrices, which are biased towards conserved sequence segments. The structural alignment does not incorporate amino acid identities and the ASTRAL40 consists of distantly related sequences only.

The fraction of sequential with respect to the non-sequential entries is at only about 7%. Therefore, further investigations must be done to accurately measure the advantage of non-sequential versus

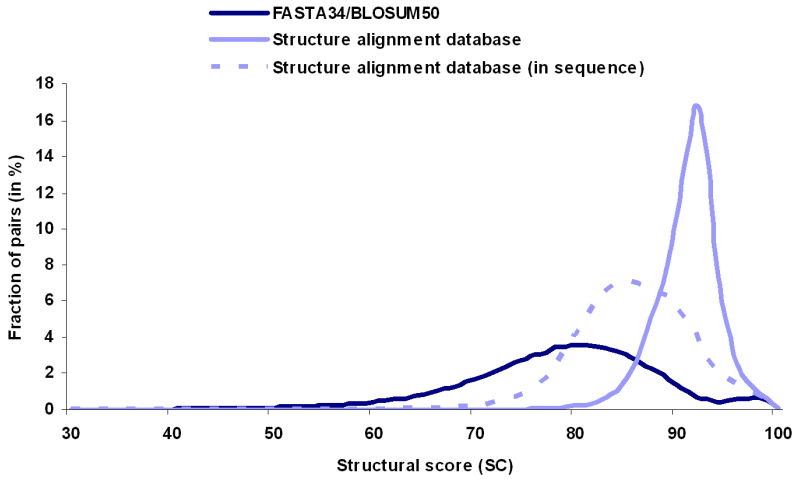


Figure 2.4. Structural score distribution for similar protein pairs with respect to sequence (dark) and structure (light). The dashed line is related to the sequential structure alignments, in which the secondary structure elements of two proteins are aligned in sequence direction.

sequential structure alignments. However, the results indicate a qualitative and quantitative gain through the non-sequential structure alignment approach. A reason for this can be the biochemical process of splicing. Furthermore, other genetic operations can reorder sequence segments. Hence, our database incorporates relations between proteins and protein families, which are less constrained by these processes. Evaluating these relations can be useful to detect alternative structures and thereby support and improve protein structure prediction methods. Further, the database can be applied as reference for other sequence based approaches.

3 DISCUSSION

The presented protein structure alignment tool GANGSTA+ solves alignment problems in a three phase hierarchical approach starting with an alignment on the secondary structure level where only α-helices and β-strands are considered. In the second phase the residue pair assignment is performed on the basis of the results from the first phase. In a subsequent last phase a refinement of the residue pair assignment is performed to complete the SSE assignment from the first phase, to find possible reassignments of SSEs and to extend the residue pair assignment beyond the SSE boundaries.

The four key features of GANGSTA+ are: (i) it can perform sequential but also non-sequential structure alignments disregarding the polypeptide connectivity; (ii) it assigns SSE pairs only if they are of the same type (α-helix, β-strand) (iii) it is capable to align SSEs having same or reverse mutual orientations; (iv) it is capable to enforce complete SSE alignments. GANGSTA+ manages to find non-sequential structure alignments, since it ignores the loops connecting the SSEs in the first SSE alignment phase. Considering the loops in the first phase already introduces a bias toward sequential SSE alignment. GANGSTA+ has the option to align SSEs in the same or opposite orientations and to enforce complete SSE alignment. Nevertheless, GANGSTA+ also aligns residues not belonging to SSEs in the third phase, if these additional residue assignments are consistent with the residue assignments within the aligned SSEs.

GANGSTA+ provides non-sequential protein structure alignments in the same time range as the fastest commonly used sequential structure alignment methods with less than a second per protein pair on average on an AMD/OPTERON with 1600MHz. Furthermore, a comparison with TM-align and the TM-score illustrates that GANGSTA+ is able to solve also sequential protein structure alignments according to the TM-score with comparable quality. We demonstrated that GANGSTA+ is able to detect non-sequential homologues for protein chains stated to possess new folds considering three examples.

In future investigations, we aim to unravel functional relationships of proteins with yet unknown functions using GANGSTA+ to detect non-sequential structure similarity. Furthermore, we aim to improve protein structure prediction approaches on the basis of non-sequential structural relations. In this context, multiple structure alignments with GANGSTA+ that can be used to define new sequence similarity measures for sequence alignment methods could be a promising direction (Schwartz and Dayhoff 1978; Pearson and Lipman. 1988; Henikoff and Henikoff 1992; Altschul et al. 1997; Pearson and Sierk 2005). Further investigations will focus on structurally similar proteins with SSE pairs aligned in reverse orientation. In contrast to the inversion of a β-strand orientation the inversion of an α-helix axis goes along with the inversion of the large helix dipole (Chakrabarti 1994). The helix dipole can have a strong influence on protein stability, its intrinsic function (Chou et al. 1988; Fairman et al. 1989; Aqvist et al. 1991; Ben-Tal and Honig 1996; D. Sengupta 2005) and ability of complex formation with other proteins (Miura et al. 1999). β-strands in proteins can be organized in alternative orientations (parallel or anti-parallel), which indicates functional robustness of proteins toward inverse orientations of β-strands. Therefore, we would expect to observe significant more β-strand inversions than α-helix inversions. GANGSTA+ enables us to analyze the functional relevance of the α-helices dipole orientation and its impact on SSE arrangements in common structural motifs for large databases. We are able to discriminate between single α-helix inversions or arbitrary many inversions of each SSE type e.g. to count parallel and anti-parallel β-strands within a certain protein family. These possibilities underline the wide range of GANGSTA+ applicability to analyze the protein fold space and its properties.

4 ADDITIONAL PUBLICATIONS

In the following, two published research papers are briefly discussed. They illustrate the research results on molecular interference of protein structures with other proteins, respectively small molecules.

4.1 Sampling geometries of protein-protein complexes

Authors Guerler A, Lorenzen S, Krull F, Knapp EW

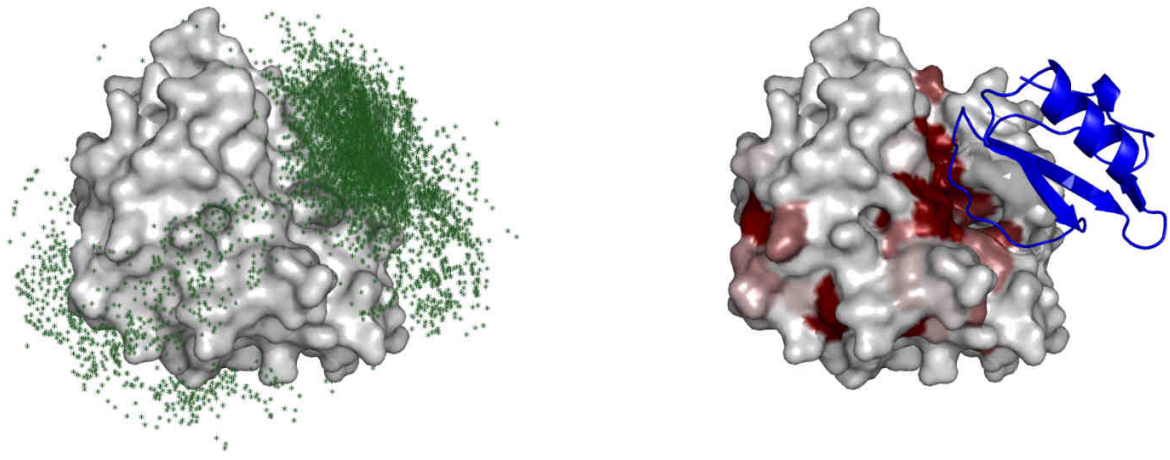
Bibliography Genome Informatics 20, 260-9, 2008

Contribution

- Development of the research question
- Development of the required software
- Generation and analysis of results
- Manuscript preparation

The full article is attached to the end of this document.

Another strategy to analyze a protein's function is to evaluate molecular interferences with other protein structures (Guerler et al. 2008a). As mentioned in the introduction, proteins are important regulators of biochemical processes in living cells. They are for instance used to catalyze chemical reactions, to transport substrates through membranes and to stabilize cellular structures. Interactions with other molecules can affect a protein's macromolecular structure and functionality. Therefore, many approaches have been developed to determine the native binding mode between two protein structures. In general, geometries of protein pairs are sampled by generating docked conformations, analyzing them with scoring functions and selecting appropriate structures for further refinement. In the following publication a fast real space algorithm to sample geometries of protein pairs is described. Initially uniformly distributed points on the surfaces of the two protein structures to be docked, and a set of uniformly distributed rotations are determined. To generate structures of protein pairs one protein of the protein pair is rotated according to a selected rotation and translated along a line connecting two surface points belonging to different proteins such that these surface points coincide. The resulting protein pair geometries are ana-



A **B**

Figure 4.1. Illustration of the docking results for the protein complex 1ACB. a) Surface of the receptor with the centers of masses of 8000 decoys (dots) with the highest score per given rotational transformation. About 10% of all decoys have an interface RMSD below 10 Å. b) Surface illustration of the receptor and cartoon illustration of the ligand molecule's secondary structure elements. The conserved residues of the receptor are highlighted in dark on the protein's surface.

lyzed and selected using an amino acid and an atom pair based contact energy function.

This is illustrated by the sampling results obtained for the first enzyme-inhibitor complex of the ZDOCK 1.0 benchmark set (see figure 4.1). 1ACB is a serine-protease-inhibitor complex (Frigerio et al. 1992). We applied the algorithm (described in detail in this publication) on separately crystallized protein structures. The surface of the serine-protease has been covered with 55 surface points and the inhibitor's structure with 23. Given the uniform set of 8000 rotations, more than 10^7 decoys have been generated. Less than 5% (387047 in total) of these decoys fulfilled the geometrical quality requirements, defined by the fraction of overlapping atoms and the normal vector deviation of assigned surface points. We calculated the interface RMSD of these decoys to the native reference complex, which was generated by aligning the separately crystallized protein structures on the co-crystallized complex structure. About 10% of them had an interface RMSD below 10 Å to the reference complex. The decoys have been scored and the highest ranked geometries per rotation kept. In 186 out of 8000 cases the protein-protein decoys exhibit a RMS deviation less than 5 Å.

In total 22 enzyme-inhibitor complexes were evaluated and the results show that an efficient sampling and scoring of unbound receptor-ligand geometries in real space is computationally feasible. The method provides decoy sets with near-native geometries for all of the considered 22 enzyme-inhibitor complexes, taking less than 30 min in total [AMD Opteron/2.2 GHz].

In future studies, we plan to utilize our method for the evaluation of a variety of other all atom, respectively heavy atom, or residue-based scoring functions. Additionally, research on new scoring schemes will be carried out. Thereby, the preliminary analysis of potential interface residues can be of particular interest.

4.2 Selection and flexible optimization of binding modes from conformation ensembles

Authors Guerler A*, Moll S, Weber M, Meyer H, Cordes F

Bibliography Elsevier BioSystems 92, 42-8, 2008

Contribution

- Development of the research question
- Development of the required software
- Generation and analysis of results
- Manuscript preparation and correspondence

http://dx.doi.org/10.1016/j.biosystems.2007.11.004

In this publication an approach is presented to elucidate the interaction of proteins with small molecules (Guerler et al. 2008b). Although this project was initiated as my Master's research project, it was refined, edited and published during the first years of my PhD research work and became relevant for this research work. The target in protein-ligand docking is to determine the binding mode of small molecules (ligands) with regard to a known protein's binding site. The protein's binding pocket describes a geometrical and physico-chemical environment. A small ligand with properties complementary to the binding pocket has a high affinity for binding. This key-lock principle is energetically driven by free energy contributions and affects the macromolecular structure of the protein, which is highly linked to the protein's functionality. To determine the binding mode, we have created the program FADO that analyzes the protein binding pocket and determines a set of energetically favorable atom coordinates (pharmasite). To model the ligand flexibility, we initially generated an ensemble of ligand conformations (Meyer et al. 2006). Then all conformations are simultaneously aligned to the coordinates of the pharmasite by a physical point matching approach. Note, that a similar point matching approach is used in the refinement phase of GANGSTA+ to transfer the SSE-level alignment to the residue-level (Guerler and Knapp

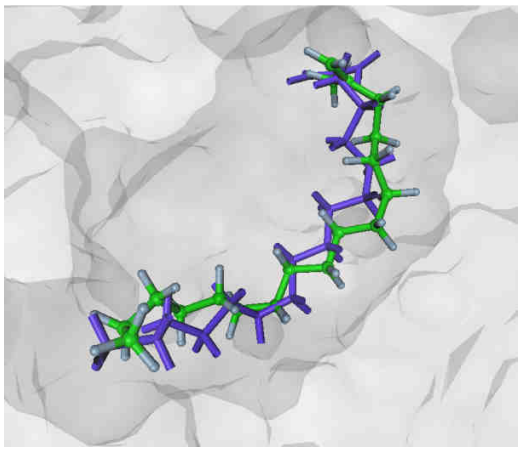


Figure 4.2. Illustration of the semi-flexible docking result for a fatty acid (green) and a transporter protein's binding pocket (gray) (2IFB, (Sacchettini et al. 1989)). The crystal reference structure is shown in blue to guide the eye and has a RMSD of 1.72 Å to the predicted binding mode.

2008). In FADO, the point matching approach, replaces the calculation of long range interactions, such as electrostatic and van-der-Waals interaction. These long range interactions are the most time consuming calculations in molecular docking. Additionally, steric repulsion with the protein is ignored, as only the pharmasite is needed for point matching. This enhances the flexibility of the ligand, not being constrained by steric repulsion during docking.

Figure 4.2 illustrates the semi-flexible docking result for a fatty acid protein complex (2IFB). The result for a benchmark dataset of 28 protein-ligand complexes is illustrated in figure 4.3. FADO reproduced 78% of the main dataset below 2Å RMSD, which is a very good performance compared to currently available docking methods. However, this takes additional 40 seconds of full flexible ligand optimization with the MMFF force field on average for each complex. Further, this only holds under the assumption that we will be able to develop an additional scoring function, which reliably selects the correct binding mode out of the 20 proposed modes currently produced by FADO. Unfortunately, the pure MMFF force field energy is not able to reliably detect the correct binding modes. Further investigations

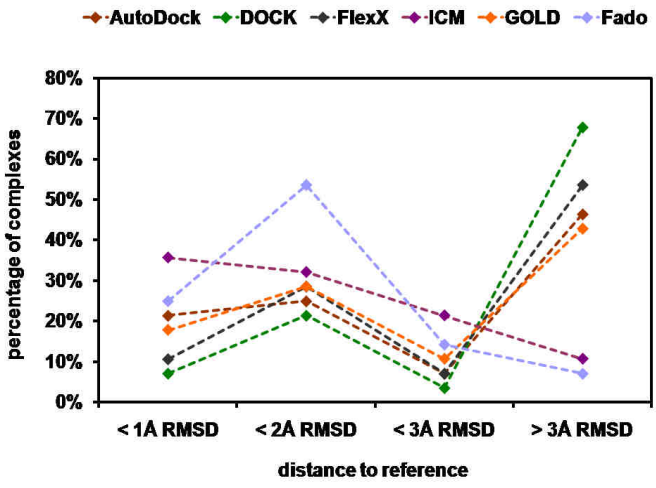


Figure 4.3. Results of flexible docking (28 complexes) with FADO compared with results of commonly used docking tools. FADO reproduced 78% of all complexes below than 2 Å RMSD (Guerler et al. 2008b).

must be done to develop a reliable and accurate scoring scheme. Alternatively, a set of common scoring functions (Böhm, 1994; Muegge, 1999; Gohlke, 2000) is available as additional Amira plugin upon request by the authors.

5 AVAILABILITY

All of the mentioned methods were written in ANSI C++ and have been tested on different platforms i.e. Linux i.e Gentoo and Ubuntu, and Windows i.e. Cygwin and MingW. GANGSTA+ is available at *http://agknapp.chemie.fu-berlin.de/gplus*, in the JAVA application STRAP (*http://www.charite.de/bioinf/strap*) maintained by Christoph Gille at the Charité Universitätsmedizin Berlin. GangstaLite is available at the 3D structural superposition server Superimposé (*http://farnsworth.charite.de/superimpose-web*) maintained by Raphael Bauer and Dr. Robert Preissner at the Charité Universitätsmedizin Berlin. FADO is available as plug-in for the molecular modeling environment AMIRA (*http://www.amiravis.com*) and is a property of the Konrad-Zuse-Institut Berlin (*http://www.zib.de*). The presented protein docking approach is available on request by the authors

6 SUMMARY IN ENGLISH

In this work, the developments and results of the non-sequential structure alignment method GANGSTA+ are presented. The method solves the structure alignment problem hierarchically. Initially the secondary structure elements of two protein structure pairs are assigned. This is achieved with a deterministic algorithm, which solves the so called contact-map-overlap problem with a combinatorial approach. After the assignment of secondary structure elements between a protein structure pair, GANGSTA+ translates the assignment to the residue-level with a point matching approach, which is described in detail in this work. On the basis of the residue-level assignment a transformation matrix is determined, which aligns the two protein structures to each other. Contrary to most other methods GANGSTA+ is able to ignore the sequential order of the assigned secondary structure elements and is even able to align secondary structure elements in sequence reversed direction. Despite the enlarged solution space, the method is approximately as fast as the fastest available sequential structure alignment methods. GANGSTA+ has been applied on several million protein structure pairs and the results have been made available to the public by online services. It could be shown that GANGSTA+ is able to find non-sequential structural analogs for protein structures stated to be novel folds. The whole functionality of GANGSTA+ is available to the public at several online services and applications.

7 ZUSAMMENFASSUNG AUF DEUTSCH

In dieser Arbeit werden die Entwicklung und die Ergebnisse der nicht-sequentiellen Proteinstrukturüberlagerungsmethode GANGSTA+ vorgestellt. Die Methode löst das Strukturüberlagerungsproblem stufenweise und bestimmt zu Anfang eine Zuordnung von Sekundärstrukturelementen zwischen einem gegebenen Proteinstrukturpaar. Dies geschieht mit einem deterministischen Algorithmus, der das sogenannte „contact-map-overlap“ Problem kombinatorisch optimiert. Auf die Zuordnung der Sekundärstrukturelemente folgt dann die Bestimmung einer eindeutigen Zuordnung der Residuen beider Proteinstrukturen. Dies wird mithilfe einer Punktüberlagerungsmethode erreicht, welche detailliert in dieser Arbeit erläutert wird. Ist die eindeutige Zuordnung auf Residuenebene bestimmt wird eine Transformationsmatrix ermittelt, welche die beiden Proteinstrukturen überlagert. Im Gegensatz zu den meisten anderen Methoden ist GANGSTA+ in der Lage die sequentielle Anordnung der Sekundärstrukturelemente zu ignorieren und sogar sequenzinvertierte strukturelle Ähnlichkeiten zu ermitteln. Trotz des größeren Lösungsraumes ist die Methode in etwa so schnell wie die schnellsten bisher entwickelten sequentiellen Strukturüberlagerungsmethoden. GANGSTA+ wurde auf mehrere Millionen Proteinpaare angewendet und die Ergebnisse sind öffentlich zugänglich gemacht worden. Es konnte gezeigt werden das GANGSTA+ in der Lage ist für Proteinstrukturen, welche als neu gelten, nicht-sequentielle Strukturanaloga zu finden. Die gesamte Funktionalität von GANGSTA steht der Öffentlichkeit über mehrere Internetseiten und Softwareanwendungen zur Verfügung.

Statutory Declaration

Hereby, I testify that this thesis is the result of my own work and research, except of references given in the bibliography. This work contains material that is the copyright property of others, which cannot be reproduced without the permission of the copyright owner. Such material is clearly identified in the text.

Aysam Guerler

References

Altschul, S.F., Madden, T.L., Schaffer, 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 Research* **25:** 3389-3402.

Anfinsen, C.B., Haber, E., Sela, M., and White, F.H. 1961. The Kinetics of Formation of Native Ribonuclease During Oxidation of the Reduced Polypeptide Chain. *Proceedings of the National Academy of Sciences* **47:** 1309-1314.

Aqvist, J., Luecke, H., Quiocho, F.A., and Warshel, A. 1991. Dipoles localized at helix termini of proteins stabilize charges. *PNAS* **88:** 2026–2030.

Bauer, R.A., Bourne, P.E., Formella, A., Frömmel, C., Gille, C., Goede, A., Guerler, A., Hoppe, A., Knapp, E.W., Pöschel, T., et al. 2008. Superimposé: a 3D structural superposition server. *Nucleic Acids Research* **36:** W47-W57.

Baumann, H., Knapp, S., Lundback, T., Ladenstein, R., and Hard, T. 1994. Solution structure and DNA-binding properties of a thermostable protein from the archaeon Sulfolobus solfataricus. *Nat.Struct.Biol.* **1:** 808-819.

Ben-Tal, N., and Honig, B. 1996. Helix-helix interactions in lipid bilayers. *Biophys J.* **71:** 3046-3050.

Benach, J., Abashidz, E.M., Jayaraman, S., Rong, X., Acton, T.B., Montelione, G.T., and Tong, L. 2005. Crystal structure of Q8ZRJ2 from salmonella typhimurium. NESG TARGET STR65. *to be published*.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E. 2000. The Protein Data Bank. *Nucleic Acids Research* **28:** 235-242.

Bowman, G.D., O`Donnell, M., and Kuriyan, J. 2004. Structural analysis of a eukaryotic sliding DNA clamp-clamp loader complex. *Nature* **429:** 724-730.

Buchner, J. 2002. The cellular protein folding machinery. *Cell Mol. Life Sci* **59:** 1587-1588.

Chakrabarti, P. 1994. An assessment of the effect of the helix dipole in protein structures. *Protein Engineering* **7:** 471-474.

Chandonia, J., Hon, G., Walker, N., Lo, C., Koehl, P., Levitt, M., and Brenner, S. 2004. The ASTRAL compendium in 2004. *Nucleic Acids Research* **32:** 189-192.

Chou, K.C., Maggiora, G.M., Némethy, G., and Scheraga, H.A. 1988. Energetics of the structure of the four-alpha-helix bundle in proteins. *PNAS* **85:** 4295-4299.

Cooper, D.N., Ball, E.V., and Krawczak, M. 1997. The human gene mutation database. *Nucleic Acids Research* **26:** 285-287.

D. Sengupta, R.B., J. Smith, G. Ullmann. 2005. The α Helix Dipole: Screened Out? *Structure* **13:** 849-855.

Fairman, R., Shoemaker, K.R., York, E.J., Stewart, J.M., and Baldwin, R.L. 1989. Further studies of the helix dipole model: effects of a free alpha-NH3+ or alpha-COO- group on helix stability. *Proteins* **5:** 1-7.

Frigerio, F., Coda, A., Pugliese, L., Lionetti, C., Menegatti, E., Amiconi, G., Schnebli, H.P., Ascenzi, P., and Bolognesi, M. 1992. Crystal and molecular structure of the bovine a-chymotrypsin-eglin c complex at 2.0 A resolution. *J. Mol. Biol.* **225:** 107-123.

Frolow, F., Kalb, A.J., and Yariv, J. 1994. Structure of a unique twofold symmetric haem-binding site. *Nat.Struct.Biol.* **1:** 453-460.

Grant, A., Lee, D., and Orengo, C. 2004. Progress towards mapping the universe of protein folds. *Genome Biology* **5**.

Guerler, A., and Knapp, E.W. 2007. Evaluation of distantly related sequence pairs with respect to structural similarity. *Genome Informatics* **18:** 183-191.

Guerler, A., and Knapp, E.W. 2008. Novel protein folds and their non-sequential structural analogs. *Protein Science* **17:** 1374-1382.

Guerler, A., Lorenzen, S., Krull, F., and Knapp, E.W. 2008a. Sampling geometries of protein complexes. *Genome Informatics* **20:** 260-169.

Guerler, A., Moll, S., Weber, M., Meyer, H., and Cordes, F. 2008b. Selection and flexible optimization of binding modes from conformation ensembles. *BioSystems* **92:** 42-48.

Guo, H.H., Choe, J., and Lawrence, L.A. 2004. Protein tolerance to random amino acid change. *PNAS* **101:** 9205-9210.

Hahn, M., Piotukh, K., Borriss, R., and Heinemann, U. 1994. Native-like in vivo folding of a circularly permuted jellyroll protein shown by crystal structure analysis. *Proc.Natl.Acad.Sci.USA* **91:** 10417-10421.

Hartley, H. 1951. Origin of the Word 'Protein'. *Nature* **168**.

He, L., and Hannon, G.J. 2004. MicroRNAs: small RNAs with a big role in gene regulation. *Nat Rev Genet* **5:** 522-531.

Henikoff, S., and Henikoff, J.G. 1992. Amino acid substitution matrices from protein blocks. *PNAS* **89:** 10915-10919.

Hill, C.P., Osslund, T.D., and Eisenberg, D. 1993. The structure of granulocyte-colony-stimulating factor and its relationship to other growth factors. *Proc.Natl.Acad.Sci.USA* **90:** 5167-5171.

Holden, H.M., Wesenberg, G., Raynes, D.A., Hartshorne, D.J., Guerriero, V., and Rayment, I. 1996. Molecular structure of a proteolytic fragment of TLP20. *Acta Crystallogr.* **52:** 1153-1160.

JB, S. 1926. The isolation and crystallization of the enzyme urease. *Journal of Biological Chemistry* **69:** 435-441.

Jr., G.N.R., Becker, J.W., and Edelman, G.M. 1975. The covalent and three-dimensional structure of concanavalin A. IV. Atomic coordinates, hydrogen bonding, and quaternary structure. *J.Biol.Chem.* **250:** 1525-1547.

Keitel, T., Meldgaard, M., and Heinemann, U. 1994. Cation binding to a Bacillus (1,3-1,4)-beta-glucanase. Geometry, affinity and effect on protein stability. *Eur.J.Biochem.* **222:** 203-214.

Kendrew, J., Bodo, G., Dintzis, H., Parrish, R., Wyckoff, H., and Phillips, D. 1958. A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. *Nature* **181:** 662–666.

Kolbeck, B., May, P., Schmidt-Goenner, T., Steinke, T., and Knapp, E.W. 2006. Connectivity independent protein-structure alignment. *BMC Bioinformatics* **7**.

Kolodny, R., Koehl, P., and Levitt, M. 2005. Comprehensive Evaluation of Protein Structure Alignment Methods. *Journal of Molecular Biology* **346:** 1173-1188.

Lodi, P.J., Ernst, J.A., Kuszewski, J., Hickman, A.B., Engelman, A., Craigie, R., Clore, G.M., and Gronenborn, A.M. 1995. Solution structure of the DNA binding domain of HIV-1 integrase. *Biochemistry* **34:** 9826-9833.

Lupas, A.N., Ponting, C.P., and Russel, R.B. 2001. On the Evolution of Protein Folds: Are Similar Motifs in Different Protein Folds the Result of Convergence, Insertion, or Relics of an Ancient Peptide World? *Journal of Structural Biology* **134:** 191-203.

Mehl, A.F., Heskett, L.D., Jain, S.S., and Demeler, B. 2003. Insights into dimerization and four-helix bundle formation found by dissection of the dimer interface of the GrpE protein from Escherichia coli. *Protein Science* **12:** 1205-1215.

Meinhart, A., and Cramer, P. 2004. Recognition of RNA polymerase II carboxy-terminal domain by 3'-RNA-processing factors. *Nature* **430:** 223-226.

Meyer, H., Moll, S., Cordes, F., and Weber, M. 2006. ConFlow - A space-based application for complete conformational analysis. *ZIB-Report* **2006**.

Miura, Y., Kimura, S., Kobayashi, S., Iwamoto, M., Imanishi, Y., and Umemura, J. 1999. Negative surface potential produced by self-assembled monolayers of helix peptides oriented vertically to a surface. *Chemical Physics Letters* **315:** 1-6.

Muirhead, H., and Perutz, M. 1963. Structure of hemoglobin. A three-dimensional fourier synthesis of reduced human hemoglobin at 5.5 Å resolution. *Nature* **199:** 633–638.

Murzin, A.G., Brenner, S.E., Hubbard, T., and Chothia, C. 1995. SCOP. *J. Mol. Biol.* **247:** 536-540.

Pearson, W.R., and Lipman., D.J. 1988. Improved Tools for Biological Sequence Comparison. *PNAS* **85:** 2444-2448.

Pearson, W.R., and Sierk, M.L. 2005. The limits of protein sequence comparison? *Curr Opin Struct Biol.* **15:** 254-260.

Rini, J.M., Hardman, K.D., Einspahr, H., Suddath, F.L., and Carver, J.P. 1993. X-ray crystal structure of a pea lectin-trimannoside complex at 2.6 A resolution. *J.Biol.Chem.* **268:** 10126-10132.

Russ, W., Lowery, D., Mishra, D., Yaffe, M., and Ranganathan, R. 2006. Natural-like function in artificial WW domains. *Nature* **437:** 579-583.

Sacchettini, J.C., Gordon, J.I., and Banaszak, L.J. 1989. Crystal structure of rat intestinal fatty-acid-binding protein. *J.Mol.Biol.* **208:** 327-339.

Schwartz, R.M., and Dayhoff, M.O. 1978. Detection of Distant Relationships Based on Point Mutation Data. *Evolution of Protein Molecules***:** 1-16.

Shulman-Peleg, A., Shatsky, M., Nussinov, R., and Wolfson, H.J. 2007. Spatial chemical conservation of hot spot interactions in protein-protein complexes. *BMC Biology* **5**.

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

Xu, G.Y., Ong, E., Gilkes, N.R., Kilburn, D.G., Muhandiram, D.R., Harris-Brandts, M., Carver, J.P., Kay, L.E., and Harvey, T.S. 1995. Solution structure of a cellulose-binding domain from Cellulomonas fimi by nuclear magnetic resonance spectroscopy. *Biochemistry* **34:** 6993-7009.