DOI: 10.1002/ijch.201200078

From Protein Structure to Function via Computational Tools and Approaches

Rachel Kolodny*[a] and Mickey Kosloff*[b]

Abstract: The three-dimensional structures of proteins are often considered fundamental for understanding their function. Yet, because of the complexity of protein structure, extracting specific functional information from structures can be a considerable challenge. Here, we present selected approaches and tools that we have developed to study and connect protein sequence, structure, and function spaces. First, we consider a *global* perspective of structure space and view the protein data bank (PDB) as a database. We highlight challenges in searching protein structure space and in using the PDB as the starting point for computational structural studies. Then we describe a *function-oriented* view and show examples of how multiple protein structures can be used to extract insights about the function and specificity of proteins at the family level.

Keywords: bioinformatics · molecular recognition · protein structures · structure-activity relationships

1. Introduction

Proteins are characterized by their amino acid sequence, their structure, and their function. A protein sequence folds into a unique structure, and similar sequences fold into similar structures. There are, however, exceptions to these rules, as detailed below. The important unit of structure is a domain – generally a single stretch of sequence (50–300 amino acids long) that interacts weakly with adjacent domains. The function of a protein is associated with one or more domains. In many ways, the three-dimensional structure of proteins has been considered the gold standard for understanding the function of a protein, yet extracting functional information from structures can be a considerable challenge.

The sequence, structure, and function of a protein are, of course, related to one another. The sequence folds into a particular three-dimensional structure, which in turn enables the protein to carry out its function. Scholars often refer to the set of all possible protein sequences as *protein sequence space*, to the set of all protein structures as *structure space*, and to *function space*.[1] These are abstract spaces, which describe different entities: sequence space – strings of letters from the 20-letter amino-acid alphabet; structure space – compact, self-avoiding chains in three-dimensional space; and function space – various definitions of molecular functions. Within these spaces, we can define different relationships among their entities (e.g., the distance between two sequences). Then, we can study the properties of these spaces and the relationships between them.

One central and determining relationship between the sequence, structure, and function of proteins is their evolution from common ancestors. To study protein evolution, we measure the similarity among the sequences, structures, and functions of current-day proteins and deduce evolutionary relationships from these measurements. Very similar sequences hint at a common ancestry. Typically, medium-sized domains are considered homologous if more than 25% of their residues are identical.[2] When examining more remote homologues, whose sequences already diverged so that their similarity is too minor to be detected by sequence alone, we often rely on significant structural and functional similarity as evidence for homology.[3] This approach assumes that the divergence of structure and function is slower than that of sequence. Thus, several studies have attempted to identify what level of sequence similarity implies structural similarity. In their 1986 seminal paper, Chothia and Lesk posed this question, and identified the correlation between sequence identity and homology mentioned above.[4] The same question was later revisited by Sander and Schneider,[5] as well as by Rost.[6] In broad strokes, all of these studies considered a set of protein pairs of

[a] R. Kolodny
Department of Computer Science
The University of Haifa
Mount Carmel, Haifa, 31905 (Israel)
e-mail: trachel@cs.haifa.ac.il

[b] M. Kosloff
Department of Human Biology
The University of Haifa
Mount Carmel, Haifa, 31905 (Israel)
e-mail: Kosloff@sci.haifa.ac.il

known sequence and structure from the PDB, and analyzed the relationship between their sequences and structural similarity. The fact that studies of this sort are based on comparisons within the PDB motivated us to develop the methods for sophisticated searches in the PDB and for the non-trivial comparisons across multiple protein structures that are detailed below.

Here, we handpicked several approaches and tools to study and connect protein sequence, structure, and function space on two separate levels – global and function-oriented. At the global level, we view the PDB as a database, and highlight some of the challenges in searching protein structure space and in using the PDB as a starting point for computational structural studies. First, we describe a basic component that underlies searching protein structure space – the comparison of two protein structures. Then, we describe characterizations of the PDB, which are important when designing a fast structural PDB search, and methods for quick and accurate searches. At the function-oriented level, we discuss how structure search and comparison can be used to predict and investigate protein function and specificity, and highlight selected approaches that show how protein structure can be used to extract insights about the function and specificity of proteins at the family level.

Mickey Kosloff is a biochemist and computational biologist at the Department of Human Biology in the University of Haifa. He earned his B.Sc. in chemistry at the Hebrew University, where he was trained as a biochemist by the late Prof. Zvi Selinger and received his M.Sc. and Ph.D. in structural and molecular biochemistry. He then sought post-doctoral training in computational biology with Prof. Barry Honig at Columbia University, followed by combined experimental and computational work with Prof. Vadim Arshavsky at Duke University. His lab focuses on deciphering how protein structure encodes interaction specificity at the family level, which, in turn, determines the connectivity of signal transduction networks. His main research activities include understanding the molecular basis for protein-protein interaction specificity among large protein families, redesigning and engineering proteins as tools to perturb and modulate signaling networks *in vivo*, and leveraging these insights and tools to address a critical need in drug design – the pinpointing of drug binding sites that take family-level specificity into account.

Rachel Kolodny is a computational biologist at the Department of Computer Science in the University of Haifa. She earned her B.Sc. and M.Sc. at the Hebrew University, working with Prof. Nati Linial and Prof. Tali Tishby in the computer science department. Then she moved to California and studied with Prof. Michael Levitt and Prof. Leonidas Guibas. Her Ph.D. from the Stanford University School of Engineering focused on how to model and compare protein structures. She then sought post-doctoral training with Prof. Barry Honig at Columbia University, where she met Dr. Kosloff, and they began their fruitful collaboration. Dr. Kolodny's research interest is the investigation of the properties of protein sequence, structure, and function spaces, and their inter-relationships (between structure and function, and between sequence and structure). In particular, she develops useful computational tools to aid in this task.

2. Protein Structural Alignment – Comparing the Geometry of Two Protein Structures

Scientists have long sought to develop tools that compare protein structures and accurately quantify their similarity. Two structures can be compared, and their similarity quantified, via a procedure called *structural alignment* – the structural analog of sequence alignment. The input to a structural alignment program is two protein structures, which can differ in size. A successful output is two matching sub-structures of equal size and similar geometry. Alternatively, a structural alignment program can report that the two structures are geometrically unrelated. In addition, the structural alignment program returns a quantitative measure of the similarity of the two equally sized sub-structures.

The challenge of structural alignment can be viewed as an optimization problem of specialized geometric scores.[7] As there is no agreed upon geometric score in the field, different methods rely on different scores, and different programs use different heuristics to optimize these scores. In general, geometric scores try to minimize the Euclidean distance between corresponding residues, after the sub-structures are optimally superimposed on one another, e.g., by minimizing the Root Mean Square Deviation (RMSD) of the sub-structures. Importantly, to avoid very short alignments (and in particular alignments of length one, which always have an RMSD of zero), geometric scores also include a component that favors long alignments. Other parameters can also be used, e.g., the number of gaps in the alignment and/or secondary structure agreement.[7,8]

Given a geometric score, quickly finding the superposition and sub-structures that optimize it is a non-trivial technical challenge. Kolodny and Linial[9] proved that the optimal solution could be found in polynomial time, for a class of scores that are amendable to the computational technique of dynamic programming. Thus, they refuted the idea that structural alignment is a non-deterministic polynomial-time-hard (NP-hard) problem. *NP-hard* describes problems that (it is believed by the computer science community) can be solved correctly only by using an impractical amount of computational time. For such problems, the only (current) course of action is using heuristics, as opposed to finding an optimal solution. Fortunately, the Kolodny and Linial study showed that *structural alignment* does not fall into this class of problems. The method used by Kolodny and Linial relies on exhaustive exploration of the space of rigid transformations, and, specifically, the exploitation of the fact that proteins reside in three-dimensional Euclidean space. However, their algorithm is far too slow for practical purposes, as its run time is proportional to the sequence lengths to the eighth power.[9]

Instead, one can use one of the many heuristic structural alignment programs, e.g., STRUCTAL,[10] CATHEDRAL,[11] CE,[12] MAMOTH,[13] Matt,[14] and SSM.[15] For reviews of structural alignment methods, see reference [16]. Given a geometric score, evaluation of the different solutions found by different programs, followed by selection of the best alignment, is straightforward. This fact makes it possible to build a *combined effort* scheme, using several structural alignment programs.[7] That is, we can use several structural alignment methods for two input structures (albeit this does require more computational time), and thereby reduce the rate of false negatives (i.e., cases in which one or more of the heuristic programs failed to identify true geometric similarity of substructures).

3. Comparing Equally Sized Protein (Sub-) Structures

A related computational task is the comparison of two protein structures or sub-structures whose sequences are identical or related. In this case, the input for the geometric comparison is two structures of the same size, N, and the alignment is trivial: the ith residues in each structure are aligned one to another, for $1 \leq i \leq N$. We emphasize that since the alignment is known, this task is very different from the one in *structural alignment*. This task is most often and routinely addressed in the Critical Assessment of Techniques for Protein Structure Prediction (CASP) experiments, which evaluate the similarity of a predicted model to an experimentally determined structure.[17]

There are several methods that measure structural similarity between two structures of equal length. The most straightforward measure is RMSD. Unfortunately, RMSD is notorious for being sensitive to outliers, which poses a particular problem in the context of CASP. Thus, other measures were developed and are also used, including GDT_TS, GDT_HA,[18] TM_Score,[19] and MaxSub.[20] GDT scores calculate the average percent of the residues in the two structures whose C-alpha atoms fall within several cutoff distances (e.g. the GDT_TS score variant uses cutoff values of 1 Å, 2 Å, 4 Å, and 8 Å). TM_Score sums $1/(1+(d/d_0)^2)$, where d is the distance between corresponding C-alpha atoms, and d_0 is a normalizing factor. The final score is then normalized by $1/L$, where L is the size of the sub-structures. For such measures, scholars identified thresholds that separate clear-cut cases of similar and non-similar structures.[5,21] Thresholds for GDT_TS can be found in reference [22] and those for TM_Score in reference [23]. Thus, one can measure the RMSD, GDT_TS, and TM_Score of two matching substructures and use these thresholds to label the aligned sub-structures as “similar” or “non-similar.”

4. Reduced Versions of the PDB

As a first step in many studies that use the PDB, researchers generate reduced sets (often referred to as non-redundant subsets) of this database. There are two important reasons to use reduced sets: (1) They are far smaller, and thus the use of *structural alignment* to compare a query structure to all the structures in the reduced set is computationally feasible (albeit still slow/computationally demanding). (2) It has been proposed that these sets are more representative of the entire set of protein structures present in the universe (hopefully correcting for the biased sampling of experimental structure determination). There are specialized programs for identifying representative sets from the PDB, notably PDBSelect,[24] and PISCES.[25] However, when using only the sequences of the proteins to generate a non-redundant version of the PDB, one makes an implicit assumption – that proteins of similar sequences have similar structures. Likewise, when predicting protein structure using homology modeling, if a template structure for modeling a target sequence is selected by sequence alone, this implicitly assumes that all sequence-similar templates are equivalent. Yet, the assumption of similar sequences implying similar structures is not always true. In particular, proteins can adopt widely different structures to accommodate the execution of a function (e.g., induced fit), or due to a changing environment (e.g., pH change).

5. Identifying Numerous Sequence Similar yet Structurally Dissimilar Pairs in the PDB

To test this assumption and to determine the extent to which sequence similarity ensures structural similarity, we[26] carried out sequence-based structural superpositions (i.e., optimal superimposition of the residues that are aligned by sequence alone) of a large number of protein pairs. We identified thousands of examples where two proteins that are similar in sequence have structures that differ significantly from one another (see Figure 3 in reference [26]). These structural differences usually have a functional basis, often relating to conformational changes that are required for the function of the proteins. The number of such identified protein pairs and the mag-

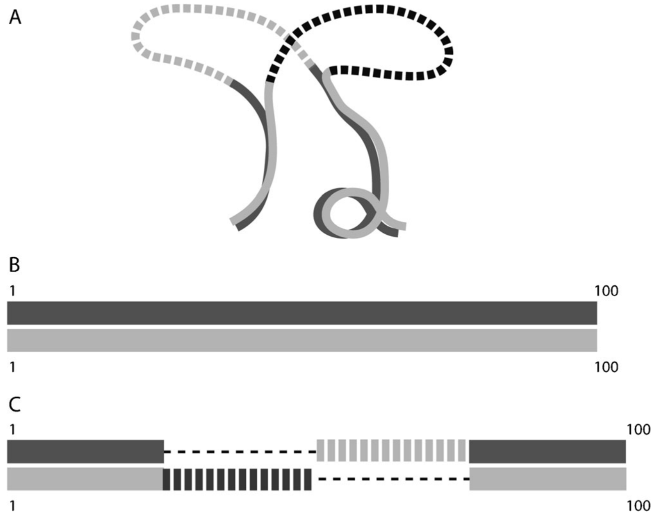

Figure 1. Structure alignment can underestimate the dissimilarity of two proteins compared to sequence-based structure superposition. To demonstrate this, we consider the case of two conformations of the same protein (i.e., the sequence identity of the two structures is 100%), superimposed one on top of the other (A). The sequence alignment of these proteins is shown in (B), and throughout the alignment the *i*th residue in the first protein matches the *i*th residue in the second protein. The RMSD calculated for the aligned matched residues will be high. On the other hand, the structural alignment of the two proteins (C) will only align the parts with similar geometry. Thus, the loops that differ will be matched with gaps. Consequently, the RMSD calculated for the residues matched in the structural alignment will be low, leading to the false conclusion that the two structures are geometrically similar.

nitude of their structural dissimilarity depend on the approach that is used to calculate the differences. In particular, we showed that the ubiquitously used geometry-based *structural alignments* will underestimate both the number of structurally dissimilar pairs and the magnitude of the structural dissimilarity (see Figure 1).

We then focused on protein pairs that share more than 99% sequence identity, yet have an RMSD greater than 6 Å. Namely, the protein pairs in this subset are essentially identical, so the structural dissimilarity cannot be attributed to low levels of sequence identity. In almost all cases, the biological function dictated a conformational plasticity that resulted in two or more distinct structures. Figure 2 lists the distribution of causes that account for the structural differences we observed for each pair in this subset. The full annotated subset is available online at (http://mt.cs.haifa.ac.il/seqsimstrdiff/seqsimstrdiff_local.htm), and includes the cause for each pair. Following this study, additional examples of protein pairs with similar sequences and non-similar structures were identified by Burra et al.,[27] and cases of conformational changes due to mutations were discussed by Murzin.[28]

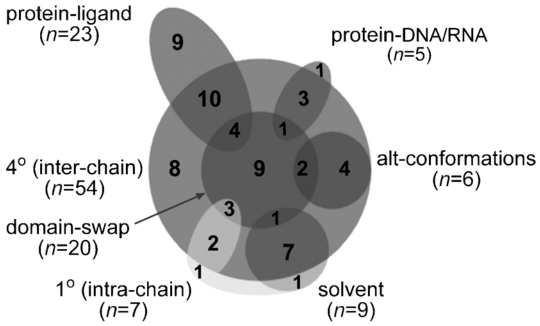

Figure 2. The Venn diagram shows the distribution of causes for the structural dissimilarity within pairs, ordered by frequency: (1) *Inter-chain (4° structure)* – different quaternary protein-protein interactions (including homomeric interactions). In the majority of cases this involves an additional protein chain, which interacts with the relevant chain in only one of the two structures in a pair. A minority of cases involved dissimilar interactions with similar binding partners (usually with an additional cause). *Domain-swap* is a subcategory of *inter-chain* interactions, where only one of the structures in a pair is domain-swapped. In rare instances both structures are domain-swapped, but with a different interface. (2) *Protein-ligand* – mostly a ligand-bound protein vs. its apo form. By "ligands," we refer to either small molecules, which are non-protein/non-nucleic acid, or short (< 15 residues) peptides. (3) *Solvent* – significant differences in the crystallization conditions (e.g. different pH or salt concentrations). (4) *Alt-conformations* – alternative crystallographic conformations of the same protein. Four of these cases were asymmetric homomers, for which *inter-chain* is an additional cause. One instance corresponded to the same protein crystallized in different space groups, and another corresponded to two alternative fits to the same crystallographic data. (5) *Intra-chain (1° structure)* – the presence/absence of part of a protein chain in one of the structures, a point mutation (combined with an additional cause), or, in two instances, oxidized vs. reduced intrachain S-S bonds. (6) *Protein-DNA/RNA* – a DNA-bound protein vs. its apo form. One instance involves a restriction enzyme (BamH) bound to specific vs. non-specific DNA sequences. *n* refers to the number of occurrences of each cause, out of the 66 separate cases examined.

6. *Filter and Refine* for Searching a Database of Protein Structures

As an alternative to relying on a reduced representative set of the PDB, scholars developed the *filter and refine* paradigm to speed up structural searches in the entire PDB (see Figure 3).[29] A *filter* method quickly sifts through a large set of structures (e.g., the complete PDB), and selects a small candidate set. Then, in the *refine* step, these candidates can be structurally aligned by a more accurate, but computationally expensive, structural alignment heuristic method. Filter methods gain their speed by representing structures abstractly – typically as vectors – and comparing these representations quickly. The vectors representing the structures in the PDB are usually calculated and stored in a pre-processing step. Then, given a query protein structure, the filter method calculates its corresponding vector and compares

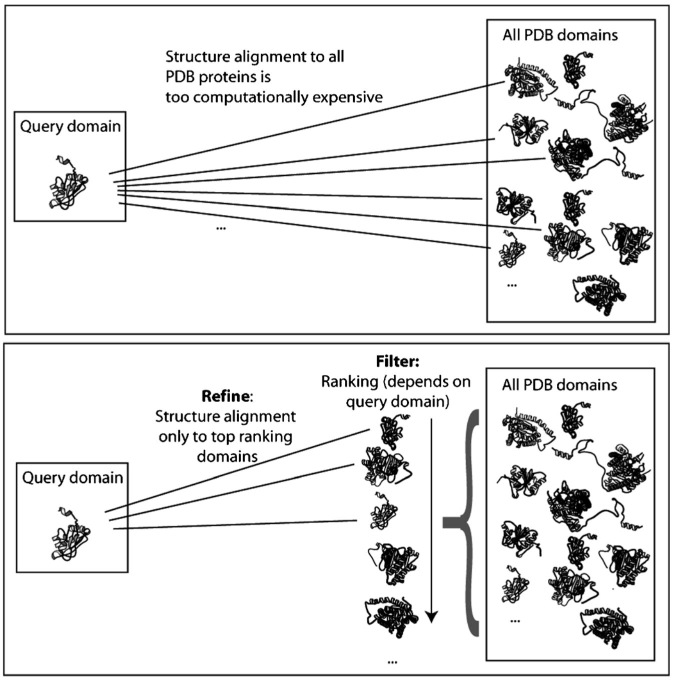

Figure 3. A schematic description of the filter and refine paradigm. The upper panel depicts a naïve search in the PDB with a query domain – compared to all domains in the PDB using structural alignment (resulting in an infeasible computation). The lower panel depicts a faster alternative: given a query domain, a fast filter step ranks the PDB domains according to how similar they are to the query. Then, in a refine step, slower but accurate structural alignment is used to compare the query domain to the top ranking domains only, thereby identifying domains in the PDB that are truly similar to the query.

it to all PDB derived vectors. Since the comparison of two vectors is a very fast computation, even a naïve comparison of all vectors, one by one, is sufficiently fast to allow structural searches against the full-sized PDB. Vector representations have an additional advantage, which holds promise for an even faster identification of similar structures: they are amendable to storage in inverted indices. An inverted index, much like a book index, is a data structure that enables fast identification and retrieval of neighbors, even in huge datasets (e.g., the index used by Google to allow fast searches of the WWW).[30]

Many different filter methods for protein structure have been developed. One such method, PRIDE, represents a structure by the histograms of diagonals in its internal distance matrix.[31] Another method, developed by Choi et al.,[32] represents a structure by a vector of frequencies of local features in its internal distance matrix. Inspired by knot theory, Rögen and Fain devised the Scaled Gauss Metric (SGM) method, which represents a structure by a vector of 30 global topological measures of its backbone.[33] Zotenko et al. represent a protein structure by a vector of the frequencies of patterns of secondary structure element (SSE) triplets.[34] There are also methods by Zhang et al.,[35] 3D-Blast,[36] YAKUSA,[37] and a method by Sacan et al.[38]

Budowski et al. presented FragBag, a filter method for identifying structurally similar domains.[39] In FragBag, each domain is represented as a fixed-size vector that describes the composition of local backbone fragments in its structure. The structural distance between two domains is approximated by the distance between their corresponding vectors. To calculate the FragBag representation of a protein domain, one needs a library of L fixed length fragments (e.g., the libraries in reference [40]). Each segment along the protein backbone is then described by its best approximation from the library of fragments. This is essentially a discretized description of the dihedral angles along the protein backbone. The FragBag vector does not record the order of fragments along the backbone. Borrowing terms from database searches in computer science, it is, therefore, a *bag*, rather than a *sequence* of fragments. Thus, the FragBag vector has L entries, and the ith entry is the number of times the ith fragment is the approximation fragment for any backbone segment (see Figure 4).

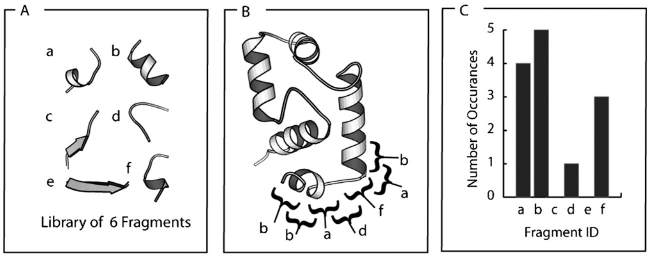

Figure 4. Description of a protein domain as a FragBag vector. As an example, we consider a library of six fragments (A). Each (overlapping) contiguous segment in the backbone of the domain is associated with its most geometrically similar library fragment (B). The structure of the protein domain is represented by a vector whose entries count the number of times each library fragment appears in this collection (C).

Budowski et al. measured how well different filter methods identify structural neighbors, and demonstrated that FragBag performs better than previous filter methods, and, surprisingly, performs comparably to computationally expensive structural alignment methods. The challenge for filter methods is ranking truly structurally similar proteins in the database high on their candidate list. Thus, to evaluate the performance of a filter method for a particular query protein structure, it is appropriate to use receiver operating characteristic (ROC) curve analysis, and to rely on a gold standard that determines which domains in the database are structural neighbors of the query. Budowski et al. used a stringent gold standard: structural neighbors found by a combined best-of-six structural alignment method. They considered the average performance over a test set of almost 3,000 CATH domains[40], and compared the rankings of FragBag filters using different fragment libraries of varying sizes and fragment lengths, and different measures of vector similarity. Then, they compared their top performer with the following rankings: (I) based on BLAST E-values, (II) previous filter methods: SGM, Zotenko et al., and PRIDE, and (III) the structural alignment methods STRUCTAL, CE, and SSM. As expected, the structural alignment methods performed best, followed by the filter methods, and then the sequence alignment method. Among the filter methods, FragBag performed best. Surprisingly, FragBag performed on par with the accurate, yet computationally expensive, structural alignment methods.

7. The Advantage of Fixed Representation of Protein Structures for Investigating the Organization of Protein Structure Space

Recently, Osadchy and Kolodny showed that the fixed-size vector representations of protein structure could also be used to draw maps of protein structure space and to investigate the relationship between protein structure and function.[41] Maps of protein structure space are visual representations of the space of all protein structures, and were previously studied by Orengo et al.,[42] Holm and Sander,[43] and by Kim and colleagues.[44] Each structure is represented by a point (in a two- or three-dimensional representation), and the distance between any two points is an approximation of the structural distance between their corresponding structures. (Of course, the structural distance depends on the particular mapping method used.) The points are then colored according to some property, e.g., the SCOP class of the protein. Such maps provide an overview of structure space that can complement hierarchical classifications such as SCOP[45] and CATH.[46]

To calculate maps of structure space, a computational procedure called Multi Dimensional Scaling (MDS) is used. MDS calculates a matrix, which maps points representing protein structures onto coordinates in three- (or two-) dimensional space. This matrix is calculated from a higher-dimensional matrix holding the structural similarities between all pairs within these structures. Importantly, calculating the all vs. all MDS matrix of many structures is a very expensive (even infeasible) computation for a large number of structures (e.g., a few thousand), and this places an effective limit on the number of protein structures in such a map.

To create maps of structure space based on a fixed-size vector representation, an equivalent, yet far faster, computational procedure can be used – Principal Component Analysis (PCA).[47] Osadchy and Kolodny used this more efficient procedure to map a very large set (>30,000) of protein domains. This allowed the study of properties such as structural density and functional diversity, which are defined at each point of structure space through the collection of structures in the vicinity of that point. The study of functional diversity is relevant for protein function prediction based on structure. By coloring the maps according to the values of these properties, their characteristics can be visualized across structure space.

The maps of the functional diversity across protein structure space revealed an unexpected relationship between structure and function: structure space has a region of high functional diversity.[41] As expected, the high functional diversity region includes the prototypic example of a multifunctional super-family – TIM barrels – but also includes many other protein folds. It consists mainly of alpha/beta structures, which are known to be the most ancient proteins.[48] The maps suggest that protein function prediction from global structure similarity is a very difficult task for structures that fall in the high functional diversity region.

One strategy for protein function prediction is to identify protein homologues of known function that have similar sequences and structures, and to transfer their functions to the target protein.[49] However, there are cases when no homologues can be identified based on sequence, and then one must resort to global structure-based function prediction. Osadchy and Kolodny[41] analyzed the relationship between the success of function predictions from global structure similarity and the location of the target protein in structure space. To do this, they relied on the dataset by Watson et al.[50] who predicted the function of 90 proteins from global structure similarity (using the structural alignment program SSM[15]) and assessed if the predictions were successful or not.

Function predictions from global structure similarity by Watson et al.[50] were more successful in regions of low functional diversity than in regions of high functional diversity. This was quantified by dividing the proteins into two sets, according to their functional diversity, and comparing the success rate in each set. The first set consists of 35 proteins in high diversity (≥ 45) vicinities, and the second consists of 55 proteins in low diversity (<45) vicinities. (The details of the measure of function diversity can be found in reference [41].) Among the high diversity proteins, only 43 % of the predictions were correct, significantly lower than the 67 % of correct predictions for the low diversity proteins ($p=0.021$ in a one-sided, two-sample proportion test). Indeed, this is also apparent from a map of successful/unsuccessful predictions within protein structure space (Figure 14S in reference [41]), in which the more successful predictions lie in the low functional diversity regions.

More generally, predicting protein function from global structure similarity is a challenging problem that is far from being solved. Therefore, functional diversity maps can be useful in providing reliable confidence measures for structure-based function predictions, and, in particular, in identifying cases where such prediction is unreliable.

8. Extracting Specific Function by Comparing Multiple Structures

In contrast to the limitation of global function prediction from structure, a focused comparison of multiple structures enables a deeper insight into function and specificity across a family of proteins. In particular, and as detailed above, comparisons of different protein structures are commonly carried out by superimposing coordinates of protein backbones. However, when the objective is analysis of similarities and differences in the active sites of different enzymes, there is an inherent problem in using the same domains for the superimposition.

To bypass this problem, Kosloff and Selinger used a comparative approach, termed *Substrate Directed SuperImposition* (SDSI). This approach entails the superimposition of multiple protein-substrate structures using the coordinates of the comparable substrates, exclusively. SDSI, therefore, provides an *unbiased* comparison of the active site environment from the substrate's point of view. In this work, SDSI was applied to various G-protein structures, in order to dissect the mechanism of the GTPase reaction that controls the signaling activity of this important family.[51] SDSI indicated that dissimilar G-proteins stabilize the transition state of the GTPase reaction in a similar fashion. This observation supported the commonality of the crucial step in this reaction – a reorientation of two critical residues, an Arginine and a Glutamine. Additionally, SDSI ascribed the catalytic inefficiency of the small G-protein Ras to the high flexibility of its active site, and downplayed the possible catalytic roles of the highly conserved Lys16 residue in Ras GTPase. This study also demonstrated that in contrast to all other Gly12 Ras mutants, which are oncogenic, the Gly12 to Pro mutant does not interfere with the catalytic orientation of the critical Glutamine. This may explain why this mutant has a higher rate of GTP hydrolysis and is non-transforming.

Another advantage of SDSI is its ability to accurately compare divergent structures that, nevertheless, bind comparable ligands. For example, SDSI reveals unexpected similarities in the divergent catalytic machineries of G-proteins and UMP/CMP kinase (Figure 5).

A somewhat different approach, which used multiple structure comparison to understand family-level specificity, looked at Glutathione S-transferases (GSTs), which comprise a diverse super-family of enzymes found in organisms from all kingdoms of life. These enzymes are involved in diverse processes, notably small-molecule biosynthesis and detoxification, and are frequently used in protein engineering studies and as biotechnology tools. Because the GST super-family is very diverse, GSTs have been subdivided into an ever-increasing number of subfamilies, or *classes*, associated with different functionalities and enzymatic specificities. This classification has usually been based on a combination of criteria, such as biochemical properties, primary, tertiary, and quaternary structure, and immunological reactivity.

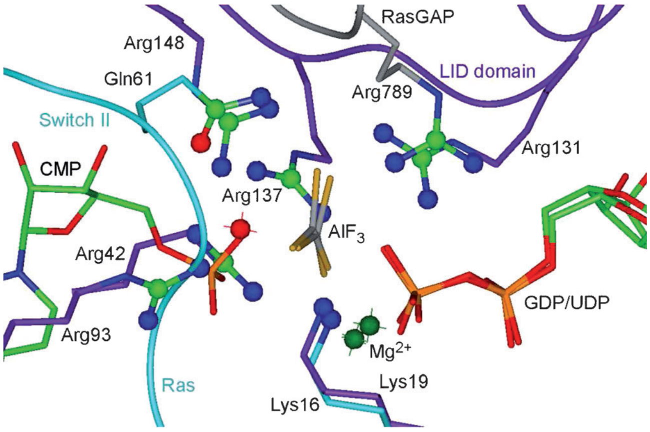

Figure 5. SDSI of the transition state structures of Ras (cyan) and UMP/CMP kinase (UMPk) (purple). A similar conformation of the P-loops (not shown) and P-loop lysines (residues 16 in Ras and 19 in UMPk) is seen in the two structures. The switch II domain in Ras and the LID and NMB binding domains (not shown) in UMPk have no correlated domains in the corresponding structure. Yet, the orientations of the functional groups of Gln61 (Ras) and Arg789 (RasGAP) relative to the substrate are highly similar to those of Arg148 and Arg131 (UMPk) respectively, suggesting a comparable catalytic role.

Through the use of a multiple structural comparison of representatives from different GST classes, Kosloff et al. identified local structural signatures that made it possible to distinguish between different GST classes (Figure 6).[52] Most of these structural signatures consist of single residues or short, but not necessarily contiguous, structural motifs. Importantly, these structural signatures have corresponding functional significances, such as differences in catalytic properties or selective dimer formation, only between members of a specific GST class. This approach allowed the classification of novel GST proteins based on structure alone, without requiring additional biochemical or immunological data. It was validated by application to the high-resolution X-ray structure of Atu5508, a putative GST from the pathogenic soil bacterium *Agrobacterium tumefaciens* (atGST1, PDB id 2FNO). This analysis suggested that atGST1 defines a new GST class, distinct from previously characterized GSTs, both in structure and in function.

Note that a central limitation to these approaches is the availability of sufficient structures of good resolution to enable such comparisons. However, the ever-increasing size of the PDB and in particular the increasing availability of multiple representatives of diverse members of large protein families, enable the application of these approaches to more biological problems.

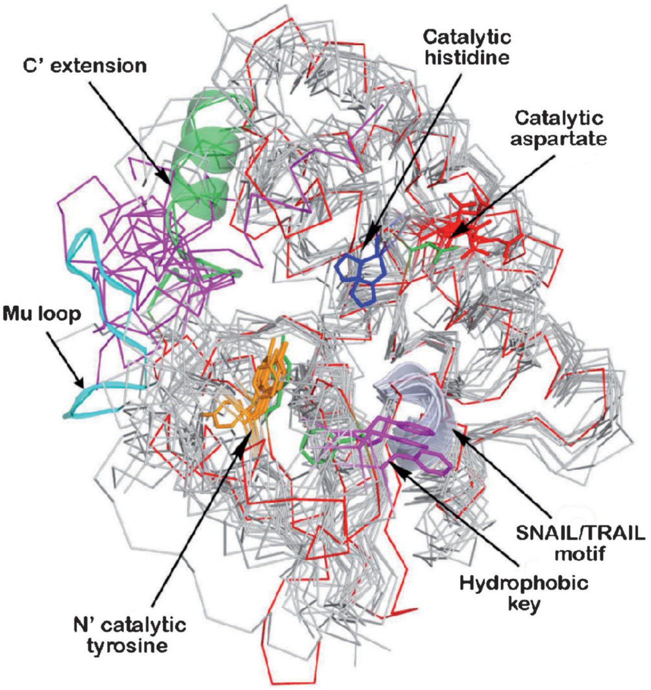

Figure 6. GST class-specific motifs shown in the context of a multiple structure alignment of representative GSTs. The Cα trace of atGST1 is colored red and all other GSTs are shown in grey. The seven motifs that define the various GST classes are labeled.

9. Deciphering Family-Level Specificity by Integrating Structure-Based Energy Calculations with Functional Assays

Intracellular signaling requires that protein-protein interactions be tailored to different signaling cascades, with either broad or narrow specificities. Understanding the basis for such selectivity is one of the major goals in signal transduction research. Yet, apart from a few representative examples, little is known of how interaction specificity is determined within large protein families. Currently, structure-based computational methods are not able to accurately predict quantitative properties, such as protein-protein binding affinities. On the other hand, while quantitative experimental comparisons offer superior accuracy, expanding such comparative approaches to an entire protein family is extremely labor intensive, and will rarely yield mechanistic insights at the resolution of individual residues.

As a model system for this problem, Kosloff et al. studied the interactions of heterotrimeric G-proteins with regulators of G-protein signaling (RGS) proteins.[53] RGS proteins function as GTPase activating proteins (GAPs) by turning G-proteins "off" and thus determining the duration of G-protein mediated signaling. The GAP mechanism of RGS proteins is particularly intriguing because, unlike other GAP proteins, RGS proteins position the catalytic machinery of G-proteins allosterically. G-proteins and RGS proteins have also been implicated in many diseases and are promising drug targets. Thus, these protein families are a major focus of both basic and applied research. Nevertheless, elucidating what determines the shared interactions or distinct specificities of these families is a difficult undertaking. Currently, computational methods are not able to predict either RGS-G-protein binding affinities or GAP activities. The alternative – an experimental comparison across these families – is a daunting task that requires testing an exorbitant number of mutants, due to the significant sequence variability among family members.

In order to understand how the structure of RGS proteins encodes their common ability to inactivate G-proteins and mediates their selective G-protein recognition, Kosloff et al. developed a new approach that integrates structure-based computations with experiments.[53] This approach combined a biochemical "benchmark" of the ability of 10 RGS domains to inactivate a G-protein with a comparative structural and energetic analysis. The latter calculated the net electrostatic/polar energetic contributions ($\Delta\Delta G_{elec}$) of each residue to the interaction with the cognate protein partner, by using a variant of *in silico* mutagenesis – perturbation of the *charges* of each residue.[54] This entailed either neutralizing a residue's backbone and side chain or neutralizing the side chain only, thereby differentiating between side-chain vs. main-chain energetic contributions. Electrostatic energies were calculated using the Finite Difference Poisson-Boltzmann method (as implemented in the DelPhi program).[55] Nonpolar energetic contributions ($\Delta\Delta G_{np}$) were calculated as a term proportional to surface-area, by multiplying the per-residue surface area buried upon complex formation (using surfv[56]) by a surface tension constant of 0.05 kcal/mol/Å.[54] Energetically significant residues were defined as those contributing $\Delta\Delta G_{elec}$ *or* $\Delta\Delta G_{np} \geq 1$ kcal/mol to the interactions.

To reduce false positives and negatives, Kosloff et al. introduced a consensus approach across the eight available structures, which substantially improved the accuracy of their predictions.[53] Residues thus determined to contribute substantially to protein-protein interaction were mapped onto a structure-to-sequence map, thereby predicting which RGS residues are essential for function and which residues can modulate specific interactions with the cognate G-protein (Figure 7). This map revealed that, in addition to previously identified conserved residues, RGS proteins contain another group of variable *modulatory residues*, which reside at the periphery of the RGS-domain/G-protein interface, and fine-tune G-protein recognition. Importantly, this residue-level map provides a shortcut that, once validated experimentally, enables the understanding of specificity, as well as its redesign, across additional family members.

These predictions were then used to redesign RGS proteins with altered function and specificity by site-specific

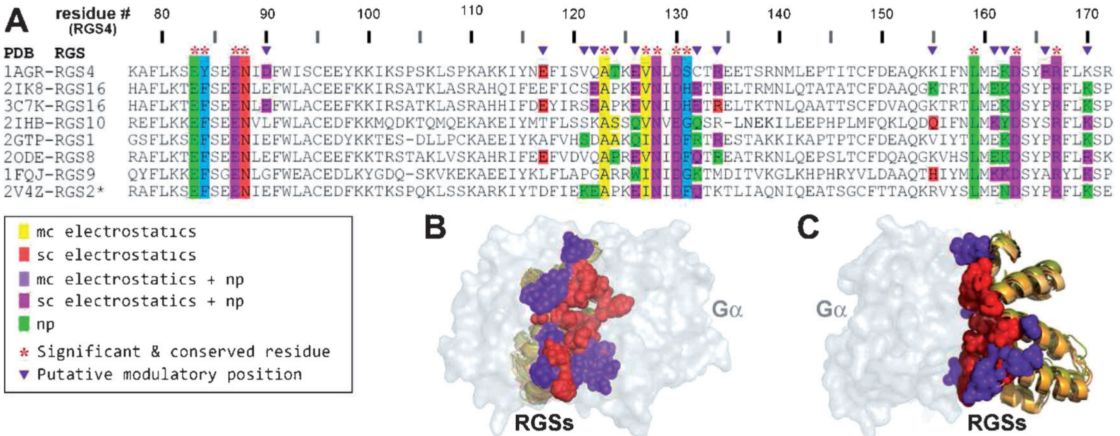
A
residue # (RGS4)
80
90
100
110
120
130
140
150
160
170
PDB RGS
1AGR-RGS4 KAFLKSEYSEENIDFWISCEEYKKIKSPSKLSPKAKKIYNEFISVQATKEVNLDSCTREETSRNMLEPTITCFDEAQKKIFNLMEKDSYRRFLKSR
2IK8-RGS16 HAFLKTEFSEENLEFWLACEEFKKIRSATKLASRAHQIFEEFICSEAPKEVNIDHETRELTRMNLQTATATCFDAAQGKTRTLMEKDSYPRFLKSP
3C7K-RGS16 HAFLKTEFSEENLEFWLACEEFKKIRSATKLASRAHHIFDEYIRSEAPKEVNIDHETRELTKTNLQAATTSCFDVAQGKTRTLMEKDSYPRFLKSP
2IHB-RGS10 REFLKKEFSEENVLFWLACEDFKKMQDKTQMQEKAKEIYMTFLSSKASSQVNVEGQSR-LNEKILEEPHPLMFQKLQDQIFNLMKYDSYSRFLKSD
2GTP-RGS1 GSFLKSEFSEENIEFWLACEDYKKTES-DLLPCKAEEIYKAFVHSDAAKQINIDFRTRESTAKKIKAPTPTCFDEAQKVIYTLMEKDSYPRFLKSD
2ODE-RGS8 RAFLKTEFSEENLEFWLACEEFKKTRSTAKLVSKAHRIFEEFVDVQAPREVNIDFQTREATRKNLQEPSLTCFDQAQGKVHSLMEKDSYPRFLKSK
1FQJ-RGS9 QYFLKKEFSGENLGFWEACEDLKYGDQ-SKVKEKAEEIYKLFLAPGARRWINIDGKTMDITVKGLKHPHRYVLDAAQTHIYMLMKKDSYARYLKSP
2V4Z-RGS2* RAFLKSEFSEENIEFWLACEDFKKTKSPQKLSSKARKIYTDFIEKEAPKEINIDFQTKTLIAQNIQEATSGCFTTAQKRVYSLMENDSYPRFLKSE
🟨 mc electrostatics
🟥 sc electrostatics
▨ mc electrostatics + np
▨ sc electrostatics + np
🟩 np
* Significant & conserved residue
▼ Putative modulatory position
B
Gα
RGSs
C
Gα
RGSs

Figure 7. Positions of significant and conserved residues and modulatory residues in multiple RGS proteins. (A) Residue-level sequence map summarizing the structure analysis and energy calculations of eight RGS–Gα crystal structures. The sequences in the multiple sequence alignment are taken from the crystal structures. RGS protein residues that contribute substantially to the interaction with Gα subunits are color-coded in the panel according to the type of their energetic contribution (see legend). Significant and conserved positions and putative modulatory positions are marked above the alignment by red asterisks and purple triangles, respectively. (B) Significant and conserved residues and modulatory residues in the eight superimposed RGS domain structures, shown as spheres, and colored red and purple, respectively. (C) Same as panel B, rotated 90° about the Y axis.

mutagenesis. Function was impaired in high-activity RGS4 and RGS16 and completely restored to low-activity RGS17 and RGS18. This approach was also applied to a completely different system – the interactions of colicin E7 with its inhibitory immunity proteins, a well-established model for studying protein-protein interaction specificity[57] – revealing novel specificity determinants.

10. Summary and Outlook

Here, we touched upon some of the challenges in using the PDB as a starting point for studying structure and function spaces, and presented some of the computational approaches we have developed for the study of protein sequence, structure, and function, and the connections among them. The database of solved proteins structures – the PDB – is rapidly increasing in size and currently holds more than 85,000 structures. In part, this increase in size is due to the high-throughput technologies for protein structure determination that have been introduced in recent years. However, these new technologies have also led to a dramatic increase in the number of proteins with known structures, yet unknown molecular functions.[58] To characterize these new structures, and more generally, to access this large dataset in a meaningful way, we need fast and accurate search methods. In this review, we used the general-purpose term "search" for (various) tasks that, given a query protein structure or sequence, allow the identification of better-studied proteins that share properties with the query protein. In particular, we focused on the important tasks of identifying and comparing proteins in the database to reveal the function of a novel protein. Ideally, search tools will be sufficiently computationally efficient to enable access to the full PDB, while returning only, or mostly, relevant results. Here, we surveyed several projects, in which we were involved, which focused on searching the PDB and relating protein structure and function spaces. We believe that characterization of the relationships among protein sequence, structure, and function spaces can be useful for developing better computational tools, and that such characterization and development of such tools are among the most important challenges facing computational structural biologists today.

References

[1] R. Kolodny, L. Pereyaslavets, A. O. Samson, M. Levitt, *Annu. Rev. Biophys.* **2013**, *42*.
[2] W. R. Pearson, *Methods Enzymol.* **1996**, *266*, 227.
[3] A. G. Murzin, *Curr. Opin. Struct. Biol.* **1998**, *8*, 380.
[4] C. Chothia, A. M. Lesk, *EMBO J.* **1986**, *5*, 823.
[5] C. Sander, R. Schneider, *Proteins: Struct., Funct., Bioinf.* **1991**, *9*, 56.
[6] B. Rost, *Protein Eng.* **1999**, *12*, 85.
[7] R. Kolodny, P. Koehl, M. Levitt, *J. Mol. Biol.* **2005**, *346*, 1173.
[8] a) L. Holm, C. Sander, *J. Mol. Biol.* **1993**, *233*, 123; b) A. S. Yang, B. Honig, *J. Mol. Biol.* **2000**, *301*, 665.
[9] R. Kolodny, N. Linial, *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 12201.
[10] S. Subbiah, D. V. Laurents, M. Levitt, *Curr. Biol.* **1993**, *3*, 141.
[11] O. C. Redfern, A. Harrison, T. Dallman, F. M. Pearl, C. A. Orengo, *PLoS Comput. Biol.* **2007**, *3*, e232.
[12] I. N. Shindyalov, P. E. Bourne, *Protein Eng.* **1998**, *11*, 739.

[13] F. Teichert, U. Bastolla, M. Porto, *BMC Bioinf.* **2007**, *8*, 425.
[14] M. Menke, B. Berger, L. Cowen, *PLoS Comput. Biol.* **2008**, *4*, e10.
[15] E. Krissinel, K. Henrick, *Acta Crystallogr., Sect. D.: Biol. Crystallogr.* **2004**, *60*, 2256.
[16] a) O. Redfern, C. Bennett, C. Orengo, in *Encyclopedia of Genetics, Genomics, Proteomics and Bioinformatics*, John Wiley & Sons, Ltd., **2004**; b) P. Koehl, *Curr. Opin. Struct. Biol.* **2001**, *11*, 348; c) M. Sierk, G. Kleywegt, *Structure* **2004**, *12*, 2103.
[17] D. Cozzetto, A. Kryshtafovych, K. Fidelis, J. Moult, B. Rost, A. Tramontano, *Proteins* **2009**, *77 Suppl 9*, 18.
[18] A. Zemla, *Nucleic Acids Res.* **2003**, *31*, 3370.
[19] Y. Zhang, J. Skolnick, *Proteins: Struct., Funct., Bioinf.* **2004**, *57*, 702.
[20] N. Siew, A. Elofsson, L. Rychlewski, D. Fischer, *Bioinformatics* **2000**, *16*, 776.
[21] B. A. Reva, A. V. Finkelstein, J. Skolnick, *Folding Des.* **1998**, *3*, 141.
[22] S. Shi, J. Pei, R. I. Sadreyev, L. N. Kinch, I. Majumdar, J. Tong, H. Cheng, B.-H. Kim, N. V. Grishin, *Database* **2009**, *2009*.
[23] a) J. Xu, Y. Zhang, *Bioinformatics* **2010**, *26*, 889; b) Y. Zhang, *Curr. Opin. Struct. Biol.* **2009**, *19*, 145.
[24] a) S. Griep, U. Hobohm, *Nucleic Acids Res.* **2010**, *38*, D318; b) U. Hobohm, M. Scharf, R. Schneider, C. Sander, *Protein Sci.* **1992**, *1*, 409.
[25] G. Wang, R. L. Dunbrack, *Bioinformatics* **2003**, *19*, 1589.
[26] M. Kosloff, R. Kolodny, *Proteins* **2008**, *71*, 891.
[27] P. V. Burra, Y. Zhang, A. Godzik, B. Stec, *Proc. Natl. Acad. Sci. U.S.A.* **2009**, *106*, 10505.
[28] A. G. Murzin, *Science* **2008**, *320*, 1725.
[29] Z. Aung, K.-L. Tan, *Drug Discovery Today* **2007**, *12*, 732.
[30] S. Brin, L. Page, *Comput. Networks ISDN* **1998**, *30*, 107.
[31] O. Carugo, S. Pongor, *J. of Mol. Biol.* **2002**, *315*, 887.
[32] I. G. Choi, J. Kwon, S. H. Kim, *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 3797.
[33] P. Rogen, B. Fain, *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 119.
[34] E. Zotenko, D. O'Leary, T. Przytycka, *BMC Struct. Biol.* **2006**, *6*, 12.
[35] Z. Zhang, H. Lee, I. Mihalek, *BMC Bioinf.* **2010**, *11*, 155.
[36] C. H. Tung, J. W. Huang, J. M. Yang, *Genome Biol.* **2007**, *8*, R31.
[37] M. Carpentier, S. Brouillet, J. Pothier, *Proteins: Struct., Funct., Bioinf.* **2005**, *61*, 137.
[38] A. Sacan, I. H. Toroslu, H. Ferhatosmanoglu, *Bioinformatics* **2008**, *24*, 2872.
[39] I. Budowski-Tal, Y. Nov, R. Kolodny, *Proc. Natl. Acad. Sci. U.S.A.* **2010**, *107*, 3481.
[40] R. Kolodny, P. Koehl, L. Guibas, M. Levitt, *J. Mol. Biol.* **2002**, *323*, 297.
[41] M. Osadchy, R. Kolodny, *Proc. Natl. Acad. Sci. U.S.A.* **2011**, *108*, 12301.
[42] C. A. Orengo, T. P. Flores, W. R. Taylor, J. M. Thornton, *Protein Eng.* **1993**, *6*, 485.
[43] L. Holm, C. Sander, *Science* **1996**, *273*, 595.
[44] a) J. Hou, G. E. Sims, C. Zhang, S. H. Kim, *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 2386; b) J. Hou, S. R. Jun, C. Zhang, S. H. Kim, *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 3651; c) I. G. Choi, S. H. Kim, *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 14056.
[45] T. J. Hubbard, B. Ailey, S. E. Brenner, A. G. Murzin, C. Chothia, *Nucleic Acids Res.* **1999**, *27*, 254.
[46] C. A. Orengo, A. D. Michie, S. Jones, D. T. Jones, M. B. Swindells, J. M. Thornton, *Structure* **1997**, *5*, 1093.
[47] J. B. Tenenbaum, V. d. Silva, J. C. Langford, *Science* **2000**, *290*, 2319.
[48] H. F. Winstanley, S. Abeln, C. M. Deane, *Bioinformatics* **2005**, *21*, 449.
[49] a) M. Punta, Y. Ofran, *PLoS Comput. Biol.* **2008**, *4*, e1000160; b) B. Rost, J. Liu, R. Nair, K. O. Wrzeszczynski, Y. Ofran, *Cell. Mol. Life Sci.* **2003**, *60*, 2637; c) A. Godzik, M. Jambon, I. Friedberg, *Cell. Mol. Life Sci.* **2007**, *64*, 2505; d) I. Friedberg, M. Jambon, A. Godzik, *Protein Sci.* **2006**, *15*, 1527.
[50] J. D. Watson, S. Sanderson, A. Ezersky, A. Savchenko, A. Edwards, C. Orengo, A. Joachimiak, R. A. Laskowski, J. M. Thornton, *J. Mol. Biol.* **2007**, *367*, 1511.
[51] M. Kosloff, Z. Selinger, *J. Mol. Biol.* **2003**, *331*, 1157.
[52] M. Kosloff, G. W. Han, S. S. Krishna, R. Schwarzenbacher, M. Fasnacht, M.-A. Elsliger, P. Abdubek, S. Agarwalla, E. Ambing, T. Astakhova, H. L. Axelrod, J. M. Canaves, D. Carlton, H.-J. Chiu, T. Clayton, M. DiDonato, L. Duan, J. Feuerhelm, C. Grittini, S. K. Grzechnik, J. Hale, E. Hampton, J. Haugen, L. Jaroszewski, K. K. Jin, H. Johnson, H. E. Klock, M. W. Knuth, E. Koesema, A. Kreusch, P. Kuhn, I. Levin, D. McMullan, M. D. Miller, A. T. Morse, K. Moy, E. Nigoghossian, L. Okach, S. Oommachen, R. Page, J. Paulsen, K. Quijano, R. Reyes, C. L. Rife, E. Sims, G. Spraggon, V. Sridhar, R. C. Stevens, H. van den Bedem, J. Velasquez, A. White, G. Wolf, Q. Xu, K. O. Hodgson, J. Wooley, A. M. Deacon, A. Godzik, S. A. Lesley, I. A. Wilson, *Proteins: Struct., Funct., Bioinf.* **2006**, *65*, 527.
[53] M. Kosloff, A. M. Travis, D. E. Bosch, D. P. Siderovski, V. Y. Arshavsky, *Nat. Struct. Mol. Biol.* **2011**, *18*, 846.
[54] F. B. Sheinerman, B. Al-Lazikani, B. Honig, *J. Mol. Biol.* **2003**, *334*, 823.
[55] B. Honig, A. Nicholls, *Science* **1995**, *268*, 1144.
[56] S. Sridharan, A. Nicholls, B. Honig, *Biophys. J.* **1992**, *6*, A174.
[57] a) G. Schreiber, A. E. Keating, *Curr. Opin. Struct. Biol.* **2011**, *21*, 50; b) . J. Mandell, T. Kortemme, *Nat. Chem. Biol.* **2009**, *5*, 797.
[58] I. Friedberg, *Briefings Bioinf.* **2006**, *7*, 225.

Received: October 7, 2012
Accepted: December 6, 2012
Published online: ■■ ■■, 0000