Sequence Order Independent Comparison of Protein Global Backbone Structures and Local Binding Surfaces for Evolutionary and Functional Inference

Joe Dundas, Bhaskar DasGupta, and Jie Liang

Abstract Alignment of protein structures can help to infer protein functions and can reveal ancient evolutionary relationship. We discuss computational methods we developed for structural alignment of both global backbones and local surfaces of proteins that do not depend on the ordering of residues in the primary sequences. The algorithm for global structural alignment is based on fragment assembly, and takes advantage of an approximation algorithm for solving the maximum weight independent set problem. We show how this algorithm can be applied to discover proteins related by complex topological rearrangement, including circularly permuted proteins as well as proteins related by complex higher order permutations. The algorithm for local surface alignment is based on solving the bi-partite graph matching problem through comparison of surface pockets and voids, such as those computed from the underlying alpha complex of the protein structure. We also describe how multiple matched surfaces can be used to automatically generate signature pockets and a basis set that represents the ensemble of conformations of protein binding surfaces with a specific biological function of binding activity. This is followed by illustrative examples of signature pockets and a basis set computed for NAD binding proteins, along with a discussion on how they can be used for discriminating NAD-binding enzymes from other enzymes.

Introduction

To understand the molecular basis of cellular processes, it is important to gain a comprehensive understanding of the biological functions of protein molecules. Although an increasing number of sequences and structures of proteins are now available, there are many proteins whose biological functions are not known, or knowledge of their biological roles is incomplete. This is evidenced by the existence

J. Liang (✉)
Bioinformatics Program, Department of Bioengineering, University of Illinois at Chicago,
Chicago, IL 60607, USA
e-mail: jliang@uic.edu

D. Kihara (ed.), *Protein Function Prediction for Omics Era*,
DOI 10.1007/978-94-007-0881-5_7, © Springer Science+Business Media B.V. 2011

of a large number of partially annotated proteins, as well as the accumulation of a large number of protein structures from structural genomics whose biological functions are not well characterized [1, 2]. Researchers have turned to in silico methods to gain biological insight into the functional roles of these uncharacterized proteins, and there has been a number of studies addressing the problem of computationally predicting the biological function of proteins [3–8].

A relatively straightforward method for inferring protein function is to transfer annotation based on homology analysis of shared characteristics between proteins. If a protein shares a high level of sequence similarity to a well characterized family of proteins, frequently the biological functions of the family can be accurately transferred onto that protein [9–11]. At lower levels of sequence similarity, probabilistic models such as profiles can be constructed using local regions of high sequence similarity [11–13]. The large amount of information of protein such as those deposited in the SWISS-PROT database [14] provides rich information for constructing such probabilistic models.

However, limitations to sequence-based homology transfer for function prediction arise when the sequence identity between a pair of proteins is less than 60% [15]. An alternative to sequence analysis is to infer protein function based on structural similarity. It is now well known that protein structures are much more conserved than protein sequences, as proteins with little sequence identity often fold into similar three-dimensional structures [16].

Protein structure and protein function are strongly correlated [17]. Conceptually, knowledge of three-dimensional structures of proteins should enable inference of protein function. Computational tools and databases for structural analysis are indispensable for establishing the relationship between protein function and structure. Among databases of protein structures, the SCOP [18] and CATH [19] databases organize protein structures hierarchically into different classes and folds based on their overall similarity in topology and fold. Such classification of protein structures generally depends on a reliable structural comparison method. Although there are several widely used methods, including Dali [20] and CE [21], current structural alignment methods cannot guarantee to give optimal results and structural alignment methods do not have the reliability and interpretability comparable to that of sequence alignment methods.

Comparing protein structures is challenging. First, it is difficult to obtain a quantitative measure of structural similarity that is generally applicable to different types of problems. Similar to sequence alignment methods, one can search for global structural similarity between overall folds or focus on local similarity between surface regions of interest. Defining a quantitative measure of similarity is not straightforward as illustrated by the variety of proposed structural alignment scoring methods [22]. Unlike sequence alignment, in which the scoring systems are largely based on evolutionary models of how protein sequence evolve [23, 24], scoring systems of structural alignment must take into account both the three-dimensional positional deviations between the aligned residues or atoms, as well as other characteristics that are biologically important. Second, many alignment methods assume the ordering of the residues follow that of the primary sequence when seeking to optimize structure similarity [21, 25]. This assumption can be problematic, as similar three-dimensional placement of residues may arise from residues with different sequential ordering. This problem is frequently encountered when comparing local regions on proteins structures. When comparing global structures of proteins, the existence of circular and higher ordered permutations [26, 27] also poses significant problems. Third, proteins may undergo minor residue side chain structural fluctuations as well as large backbone conformational changes in vivo. These structural fluctuations are not represented in a static snapshot of a crystallized structures in the Protein Data Bank (PDB) [28]. Many structural alignment methods, which assume rigid bodies and cannot account for structural changes that may occur.

In this chapter, we will first discuss several overall issues important for protein structural alignment. We then discuss a method we have developed for sequence order independent structural alignment at both the global and local level of protein structure. This is followed by discussion on how this method can be used to detect protein pairs that appear to be related by simple and complex backbone permutations. We will then describe the use of local structural alignment in automatic construction of *signature pockets* of binding surfaces, which can be used to construct *basis set* for a specific biological function. These constructs can detect structurally conserved surface regions and can be used to improve the accuracy of protein function prediction.

Structural Alignment

Protein structural alignment is an important problem [22]. It is particularly useful when comparing two proteins with low sequence identity between them. A widely used measure of protein structural similarity is the root mean squared distance (RMSD) between the equivalent atoms or residues of the two proteins. When the equivalence relationship between structural elements are known, a superposition described by a rotation matrix R and a translation vector T that minimizes the root mean squared distances (RMSD) between the two proteins can be found by solving the minimization problem:

$$\min \sum_{i=1}^{N_B} \sum_{j=1}^{N_A} |T + RB_i - A_j|^2, \qquad (1)$$

where N_A is the number of points in structure A and N_B is the number of points in structure B and it is assumed that $N_A = N_B$. The least-squares estimation of the transformation parameters R and T in Eq. (1) can be found using the technique of singular value decomposition [29].

However, it is often the case that the equivalences between the structural elements are not known a priori. For example, when two proteins have diverged significantly. In this case, one must use heuristics to determine the equivalence relationship, and the problem of protein structural alignment becomes a multi-objective problem. That is, we are interested in finding the maximum number of equivalent elements as well as in minimizing the RMSD upon superposition of the equivalent elements of the two proteins.

A number of methods that are heuristic in nature have been developed for aligning protein structures [30–37]. These methods can be divided into two categories. *Global* structural alignment methods, which are suited for detecting similarities between the overall backbones of two proteins, while *local* structural alignment methods are suited for detecting similarities between local regions or sub-structures within the two proteins. As discussed earlier, many structural alignment algorithms are constrained to find only structural similarities where the order of the structural elements follows their order in the primary sequence. Sequence order independent methods ignore the sequential ordering of the structural elements and are better suited to find more complex global structural similarities. They are also very effective for all atom comparison of protein sub-structures, as in the case of binding surface alignment. Below we discuss methods for both global and local sequence order independent structural alignment.

Global Sequence Order Independent Structural Alignment

Global sequence order independent structural alignment is a powerful tool that can be used to detect similarities between two proteins that have complex topological rearrangements, including permuted structures. Permuted proteins can be described as two proteins with similar three-dimensional spatial arrangement of secondary structures, but with a different backbone connection topology. An example of permuted proteins are proteins with circular permutations, which can be thought of as ligation of the N- and C-termini of a protein, and cleavage somewhere else on the protein. Circular permutations are interesting not only because they tend to have similar three-dimensional structure but also because they often maintain the same biological function [26]. Circularly permuted proteins may provide a generic mechanism for introducing protein diversity that is widely used in evolution. Detecting circular permutations is also important for homology modeling, for studying protein folding, and for designing protein.

A Fragment Assembly Based Approach to Sequence Order Independent Structural Alignment

We have developed a sequence order independent structural alignment method that is well-suited for detecting circular permutation as well as more complex topological rearrangement relationships among proteins [27]. Our algorithm is capable of aligning two protein backbone structures independent of the secondary structure element connectivity. Briefly, the two proteins to be aligned are first separately and exhaustively fragmented. Each fragment $\lambda^A_{i,k}$ from protein structure S_A is then pair-wise superimposed onto each fragment $\lambda^B_{j,k}$ from protein structure S_B, forming a set of fragment pairs $\chi_{i,j,k}$, where $i \in S_A$ and $j \in S_B$ are the indices in the primary sequence of the first residue of the two fragment, respectively. Here

$k \in \{5, 6, 7\}$ is the length of the fragment. For each fragment, we assign a similarity score,

$$\sigma(\chi_{i,j,k}) = \alpha \left[C - s(\chi_{i,j,k}) \cdot \frac{cRMSD}{k^2} \right] + SCS, \tag{2}$$

where *cRMSD* is the measured RMSD value after optimal superposition of the two fragments, α and C are two constants, $s(\chi_{i,j,k})$ is a scaling factor to the measured RMSD values that depends on the secondary structure of this fragment, and *SCS* is a BLOSSUM-like measure of similarity in sequence of the matched fragments [24]. Details of the similarity score and the parameters α and C can be found in [27].

The goal of structural alignment for the moment seeks to find a consistent set of fragment pairs $\Delta = \{\chi_{i_1,j_1,k_1}, \chi_{i_2,j_2,k_2}, \ldots, \chi_{i_t,j_t,k_t}\}$ that minimize the global RMSD. Finding the optimal combination of fragment pairs is a special case of the well known maximum weight independent set problem in graph theory. This problem is MAX-SNP-hard. We employ an approximation algorithm that was originally described for scheduling split-interval graphs [38] and is itself based on a fractional version of the local-ratio approach.

Our method begins by creating a conflict graph $G = (V, E)$, where a vertex is defined for each aligned fragment pair. Two vertices are connected by an edge if any of the fragments $\left(\lambda^A_{i,k}, \lambda^A_{i',k'}\right)$ or $\left(\lambda^B_{j,k}, \lambda^B_{j','k'}\right)$ from the aligned pair is not disjoint, that is, if both fragments from the same protein share one or more residues. For each vertex representing aligned fragment pair, we assign three indicator variables x_χ, $y_{\chi_{\lambda_A}}$, and $y_{\chi_{\lambda_B}} \in \{0, 1\}$ and a closed neighborhood $\mathrm{Nbr}[\chi]$. x_χ indicates whether the fragment pair should be used ($x_\chi = 1$) or not ($x_\chi = 0$) in the final alignment. $y_{\chi_{\lambda_A}}$, and $y_{\chi_{\lambda_B}}$ are artificial indicator values for λ_A and λ_B, which allow us to encode consistency in the selected fragments. The closed neighborhood of a vertex χ of G is $\{\chi' | \{\chi, \chi'\} \in E\} \cup \{\chi\}$, which is simply χ and all vertices that are connected to χ by and edge.

Our algorithm for sequence order independent structural alignment can now be described as follows. To begin, we initialize the structural alignment Δ equal to the entire set of aligned fragment pairs. We then:

1. Solve a linear programming (LP) formulation of the problem:
   maximize

$$\sum_{\chi \in \Delta} \sigma(\chi) \cdot x_\chi \tag{3}$$

   subject to

$$\sum_{a_t \in \lambda^A} y_{\chi_{\lambda_A}} \quad \leq 1 \quad \forall a_t \in S_A \tag{4}$$

$$\sum_{b_t \in \lambda^B} y_{\chi_{\lambda_B}} \quad \leq 1 \quad \forall b_t \in S_B \tag{5}$$

$$y_{\chi\lambda_A} - x_\chi \geq 0 \quad \forall \chi \in \Delta \tag{6}$$

$$y_{\chi\lambda_B} - x_\chi \geq 0 \quad \forall \chi \in \Delta \tag{7}$$

$$x_\chi, y_{\chi\lambda_A}, y_{\chi\lambda_B} \geq 0 \quad \forall \chi \in \Delta \tag{8}$$

2. For every vertex $\chi \in V_\Delta$ of G_Δ, compute its *local conflict number* $\alpha_\chi = \sum_{\chi' \in \mathrm{Nbr}_\Delta[\chi]} x_{\chi'}$. Let $\chi_{\min}$ be the vertex with the *minimum* local conflict number. Define a new similarity function σ_{new} from σ as follows:

$$\sigma_{\text{new}}(\chi) = \begin{cases} \sigma(\chi), & \text{if} \quad \chi \notin \mathrm{Nbr}_\Delta[\chi_{\min}] \\ \sigma(\chi) - \sigma(\chi_{\min}), & \text{otherwise} \end{cases}$$

3. Create $\Delta_{\text{new}} \subseteq \Delta$ by removing from Δ every substructure pair χ such that $\sigma_{\text{new}}(\chi) \leq 0$. Push each removed substructure on to a stack in arbitrary order.
4. If $\Delta_{\text{new}} \neq \emptyset$ then repeat from step 1, setting $\Delta = \Delta_{\text{new}}$ and $\sigma = \sigma_{\text{new}}$. Otherwise, continue to step 5.
5. Repeatedly pop the stack, adding the substructure pair to the alignment as long as the following conditions are met:

   a. The substructure pair is consistent with all other substructure pairs that already exist in the selection.
   b. The *cRMSD* of the alignment does not change beyond a threshold. This condition bridges the gap between optimizing a local similarity between substructures and optimizing the tertiary similarity of the alignment. It guarantes that each substructure from a substructure pair is in the same spatial arrangement in the global alignment.

Detecting Permuted Proteins

This algorithm is used in a large scale study, where a subset with 3,336 protein structures taken from the PDBSELECT 90 data set % [39] are structurally aligned in a pair-wise fashion. Our goal is to determine if we could detect structural similarities with complex topological rearrangements such as circular permutations. From this subset of 3,336 proteins, we aligned two proteins if they met the following conditions: the difference in their lengths was no more than 75 residues, and they had approximately the same secondary structure content (see [27] for details). Within the approximately 200,000 alignments, we found many known circular permutations, and three novel circular permutations previously unknown, as well as a pair of non-cyclic complex permuted proteins. Below we describe in some details the circular permutations we found between a neucleoplasmin-core and an auxin binding protein, as well as details of the more complex non-cyclic permutation.

Nucleoplasmin-Core and Auxin Binding Protein

A novel circular permutation was detected between the nucleoplasmin-core protein in *Xenopu laevis* (PDB ID `1k5j`, chain E) [40] and the auxin binding protein in maize (PDB ID `1lrh`, chain A, residues 37 through 127) [41]. The structural alignment between `1k5jE` (Fig. 1a, top) and `1lrhA` (Fig. 1a, bottom) consisted of 68 equivalent residues superimposed with an RMSD of 1.36 Å. This alignment is statistically significant with a *p*-value of 2.7×10^{-5} after Bonferroni correction. Details of *p*-value calculation can be found in reference [27]. The short loop connecting two antiparallel strands in nucleoplasmin-core protein (in circle, top of Fig. 1b) becomes disconnected in auxin binding protein 1 (in circle, bottom of Fig. 1b), and the N- and C- termini of the nucleoplasmin-core protein (in square, top of Fig. 1b) are connected in auxin binding protein 1 (square, bottom of Fig. 1b). For details of other circular permutations we discovered, including permutations between aspartate racemase and type II 3-dehydrogenase and between microphage migration inhibition factor and the C-terminal domain of arginine repressor, please see [27].

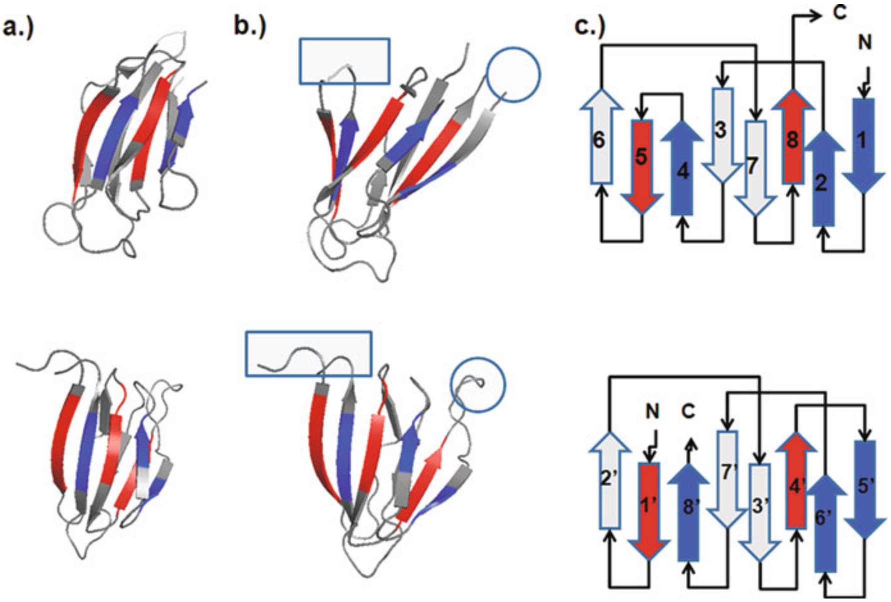

Fig. 1 A newly discovered circular permutation between nucleoplasmin-core (`1k5j`, chain E, *top panel*), and a fragment of auxin binding protein 1 (residues 37–127) (`1lrh`, chain A, *bottom panel*). **a** These two proteins align well with a RMSD value of 1.36 Å over 68 residues, with a significant *p*-value of 2.7×10^{-5} after Bonferroni correction. **b** The loop connecting strand 4 and strand 5 of nucleoplasmin-core (in *rectangle*, *top*) becomes disconnected in auxin binding protein 1. The N- and C- termini of nucleoplasmin-core (in *rectangle*, *top*) become connected in auxin binding protein 1 (in *rectangle*, *bottom*). To aide in visualization of the circular permutation, residues in the N-to-C direction before the cut in the nucleoplasmin-core protein are colored *red*, and residues after the cut are colored *blue*. **c** The topology diagram of these two proteins. In the original structure of nucleoplasmin-core, the electron density of the loop connecting strand 4 and strand 5 is missing in the PDB structure file. This figure is modified from [27]

Beyond Circular Permutation

Because of its relevance in understanding the functional and folding mechanism of proteins, circular permutations have received much attention [28, 42]. A more challenging class of permuted proteins is that of the non-cyclic permutation with more complex topological changes. Very little is known about this class of permuted proteins, and the detection of non-cyclic permutations is challenging task [43–46].

Non-cyclic permutations of the Arc repressor were created artificially were found to be thermodynamically stable. It can refold on the sub-millisecond time scale, and can bind operator DNA with nanomolar affinity [47], indicating that naturally occurring non-cyclic permutations may be as rich as the cyclic permutations. Our database search uncovered a naturally occurring non-cyclic permutation between chain F of AML1/Core Binding Factor (AML1/CBF, PDB ID `1e50`, Fig. 2a, top) and chain A of riboflavin synthase (PDB ID `1pkv`, Fig. 2a, bottom) [48, 49]. The

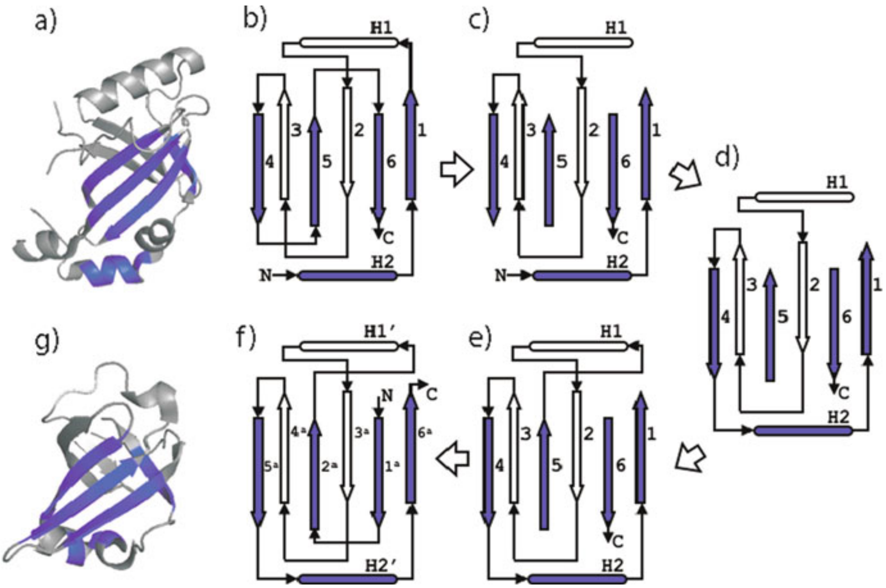

Fig. 2 A non-cyclic permutation discovered between AML1/Core Binding Factor (AML1/CBF, PDB ID `1e50`, Chain F, *top*) and riboflavin synthase (PDBID `1pkv`, chain A, *bottom*) **a** These two proteins structurally align with an RMSD of 1.23 Å over 42 residues , and has a significant *p*-value of 2.8×10^{-4} after Bonferroni correction. The residues that were assigned equivalences from the structural alignment are colored blue. **b** These proteins are related by a complex permutation. The steps to transform the topology of AML1/CBF (*top*) to riboflavin (*bottom*) are as follows: **c** Remove the loops connecting strand 1 to helix 2, strand 4 to strand 5, and strand 5 to helix 6; **d** Connect the C-terminal end of strand 4 to the original N-termini; **e** Connect the C-terminal end of strand 5 to the N-terminal end of helix 2; **f** Connect the original C-termini to the N-terminal end of strand 5. The N-terminal end of strand 6 becomes the new N-termini and the C-terminal end of strand 1 becomes the new C-termini. We now have the topology diagram of riboflavin synthase. This figure was modified from [27]

two structures align well with an RMSD of 1.23 Å, at an alignment length of 42 residues, with a significant p-value of 2.8×10^{-4} after Bonferroni correction.

The topology diagram of AML1/CBF (Fig. 2b) can be transformed into that of riboflavin synthase (Fig. 2f) by the following steps: Remove the loops connecting strand 1 to helix 2, strand 4 to strand 5, and strand 5 to strand 6 (Fig. 2c). Connect the C-terminal end of strand 4 to the original N-termini (Fig. 2d). Connect the C-terminal end of strand 5 to the N-terminal end of helix 2 (Fig. 2e). Connect the original C-termini to the N-terminal end of strand 5. The N-terminal end of strand 6 becomes the new N-termini and the C-terminal end of strand 1 becomes the new C-termini (Fig. 2f).

Local Sequence Order Independent Structural Alignment

The comparison of overall structural folds regardless of topological reconnections can lead to insight into distant evolutionary relationship. However, similarity in overall fold is not a reliable indicator of similar function [50–52]. Several studies suggest that structural similarities between local surface regions where biological function occurs, such as substrate binding sites, are a better predictor of shared biological function [8, 53–57].

Substrate binding usually occurs at concave surface regions, commonly referred to as *surface pockets* [55, 58–60]. A typical protein has many surface pockets, but only a few of them present a specific three-dimensional arrangement of chemical properties conducive to the binding of a substrate. This protein must maintain this physiochemical environment throughout evolution in order to maintain its biological function. For this reason, shared structural similarities between *functional surfaces* among proteins may be a strong indicator of shared biological function. This has lead to a number of promising studies, in which protein functions can be inferred by similarity comparison of local binding surfaces [55, 61–64].

A challenging problem with the structural comparison of protein pockets lies in the inherent flexibility of the protein structure. A protein is not a static structure represented by a Protein Data Bank entry. The whole protein as well as the local functional surface may undergo large structural fluctuations. The use of a single surface pocket structure as a representative template for a specific protein function will often result in many false negatives. This is due to the inability of a single representative to capture the full functional characteristics across all conformations of the protein.

To address this problem, we have developed a method that can automatically identify the structurally preserved atoms across a family of protein structures that are functionally related. Based on sequence-order independent surface alignments across the functional pockets of a family of protein structure, our method creates *signature pockets* by identifying structurally conserved atoms and measuring their fluctuations. As more than one signature pocket may result for a single functional class, the signature pockets can be organized into a *basis set* of signature pockets for that functional family. These signature pockets of the binding surfaces then can be used for scanning a protein structure database for function inference.

Bi-partite Graph Matching Approach to Structural Alignment

Our method for surface alignment is sequence order independent. It is based on a maximum weight bi-partite graph matching formulation of [65] with further modifications. This alignment method is a two step iterative process. First, an optimal set of equivalent atoms under the current superposition are found using a bi-partite graph representation. Second, a new superposition of the two proteins is determined using the new equivalent atoms from the previous step. The two steps are repeated until a stopping condition has been met.

To establish the equivalence relationship, two protein functional pocket surfaces S_A and S_B are represented as a graph, in which a node on the graph represent an atom from one of the two functional pockets. The graph is bi-partite if edges only connect nodes from protein S_A to nodes from protein S_B. In our implementation, directed edges are only drawn from nodes of S_A to nodes of S_B if a similarity threshold is met. The similarity threshold used in our implementation is a function of spatial distances and chemical differences between the corresponding atoms (see [66] for details). Each edge $e_{i,j}$ connecting node i to node j is assigned a weight $w(i,j)$ equal to the similarity score between the two corresponding atoms. A set of equivalence relations between atoms of S_A and atoms of S_B can be found by selecting a subset of the edges connecting nodes of S_A to S_B, with maximized total edge weight, where at most one edge can be selected for each atom [67]. A solution to the maximum weight bi-partite graph matching problem can be found using the Hungarian algorithm [68].

The Hungarian method works as follows. To begin, an overall score $F_{\text{all}} = 0$ is initialized, and an artificial source node s and an artificial destination node d are added to the bi-partite graph. Directed edges with 0-weight from the source node s to each node of S_A and from each node of S_B to the destination node d are also added. The algorithm then proceeds as follows:

1. Find the shortest distance $F(i)$ from the source node s to every other node i using the Bellman-Ford [69] algorithm.
2. Assign a new weight $w'(i,j)$ to each edge that does not originate from the source node s as follows,

$$w'(i,j) = w(i,j) + [F(i) - F(j)]. \tag{9}$$

3. Update F_{all} as $F_{\text{all}}{}' = F_{\text{all}} - F(d)$
4. Reverse the direction of the edges along the shortest path from s to d.
5. If $F_{\text{all}} > F(d)$ and a path exists between s and d then start again at step 1.

The Hungarian algorithm terminates when either there is no path from s to d or when the shortest distance from the source node to the destination node $F(d)$ is greater than the current overall score F_{all}. The bi-partite graph will now consist of directed edges that have been reversed (point from nodes of S_B to nodes of S_A). These flipped edges represent the current equivalence relationships between atoms of S_A and atoms of S_B.

The equivalence relations can then be used to superimpose the two proteins. After superposition, a new bi-partite graph is created and the maximum weight bi-partite matching algorithm is called again. This process is repeated iteratively until the change in RMSD upon superposition falls below a threshold.

Signature Pockets and Basis Set of Binding Surface for a Functional Family of Proteins

Based on the pocket surface alignment algorithm, we have developed a method that automatically generate structural templates of local surfaces, called *signature pockets*, which can be used to represent an enzyme function or a binding activity. These signature pockets contain broad structural information as well as discriminating ability.

A signature pocket is derived from an optimal alignment of precomputed surface pockets in a sequence-order-independent fashion, in which atoms and residues are aligned based on their spatial correspondence when maximal similarity is obtained, regardless how they are ordered in the underlying primary sequences. Our method does not require the atoms of the signature pocket to be present in all member structures. Instead, signature pockets can be created at varying degrees of partial structural similarity, and can be organized hierarchically at different level of binding surface similarity.

The input to the signature pocket algorithm is a set of functional pockets from a pre-calculated database of surface pockets and voids on proteins, such as those contained in the CASTp database [60]. The algorithms begins by performing all vs all pair-wise sequence order independent structural alignment on the input functional surface pockets. A distance score, which is a function of the RMSD and the chemistry of the paired atoms from the structural alignment, is recorded for each aligned pair of functional pockets (see [66] for details). The resulting distance matrix is then used by an agglomerative clustering method, which generates a hierarchical tree. The signature of the functional pockets can then be computed using a recursive process following the hierarchical tree.

The process begins by finding the two closest siblings (pockets S_A and S_B), and combining them into a single surface pocket structure S_{AB}. Because of the recursive nature of this algorithm, either of the two structures being combined may themselves already be a combination of several structures. When combining the two structures, we follow the criteria listed below:

1. If two atoms were considered equivalent in a structural alignment, a single coordinate is created in the new structure to represent both atoms. The new coordinate is calculated by averaging the coordinates of all underlying atoms that are currently represented by the two coordinates to be averaged.
2. If no equivalence was found for an atom during the structural alignment, the coordinates of that atom are transferred directly into the new pocket structure.

During each step in combining two surface pockets, a count of the number of times that an atom at the position i was present in the underlying set of pockets is recorded, which is then divided by the number of the constituent pockets. This is the *preservation ratio* $\rho(i)$. In addition, the mean distance of the coordinates of the aligned atoms to their geometric center is recorded as the *location variation* v. At the end of each step, the new structure S_{AB} replaces the two structures S_A and S_B in the hierarchical tree, and the process is repeated on the updated hierarchical tree. At a specific height of the hierarchical tree, different signature pockets can be created with different extents of structural preservation by selecting a similarity threshold value.

The signature pocket algorithm can be terminated at any point during its traversal of the hierarchical tree. Figure 3 illustrates this point by showing three different stopping thresholds (horizontal dashed lines). Depending on the choice of the threshold, one or multiple signature pockets may result. Figure 3a shows a low threshold which results in a set of 3 signature pockets. Raising the threshold can produce fewer signature pockets (Fig. 3b). A single signature pocket that represents all surface pockets in the data set can be generated by raising the threshold even further (Fig. 3c). Since clusters from the hierarchical tree represent a set of surface pockets that are similar within certain threshold, if a stopping threshold is chosen such that there exist multiple clusters in the hierarchical tree, a signature pocket will be created for each cluster. The set of signature pockets from different clusters collectively form a *basis set* of signature pockets, which represent the ensemble of differently sampled conformations for a functional family of proteins. As a basis set of signatures can represent many possible variations in shapes and chemical textures, it can represent structural features of an enzyme function with complex binding activities, and can also be used to accurately predict enzymes function.

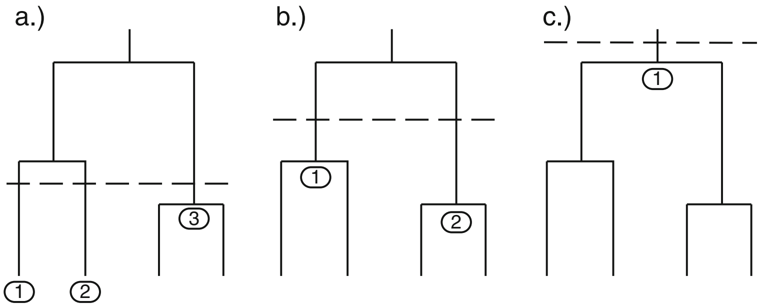

Fig. 3 Different basis sets of signature pockets can be produced at different levels of structural similarity by raising or lowering the similarity threshold (*vertical dashed line*). **a** A low threshold will produce more signature pockets. **b** As the threshold is raised, fewer signature pockets will be created. **c** A single signature pocket can in principle be created to represent the full surface pocket data set by raising the threshold

Signature Pockets of NAD Binding Proteins

To illustrate how signature pockets and basis set help to identify key structural elements important for binding and how they can facilitate function inference, we discuss a study of the nicotinamide adenine dinucucleotide (NAD) binding proteins. NAD consists of two nucleotides, nicotinamide and adenine, which are joined by two phosphate groups. NAD plays essential roles in metabolism where it acts as a coenzyme in redox reactions, including glycolysis and the citric acid cycle.

Using a set of 457 NAD binding proteins of diverse fold structures and diverse evolutionary origin, we first extracted the NAD binding surfaces from precomputed CASTp database of protein pockets and voids [60]. Based on similarity values from a comprehensive all-against-all sequence order independent surface alignment, we obtain a hierarchical tree of NAD binding surfaces. The resulting 9 signature pockets of the NAD binding pocket form a basis set, which are shown in Fig. 4.

These signature pockets contain rich biological information. Among the NAD-binding oxioreductase, three signature pockets (Fig. 4e, h, and i) are for clusters of oxioreductases that act on the CH-OH group of donors (alcohol oxioreductases), one signature pocket (Fig. 4j) is for a cluster that act on the aldehyde group of donors, and the remaining two signature pockets (Fig. 4f and g) are for oxioreductases that act on the CH-CH group of donors. For NAD-binding lyase, one of the two signature pockets (Fig. 4d) represent lyase that cleave both C–O and P–O bonds. The other signature pocket (Fig. 4b) represent lyases that cleave both C–O and C–C bonds. These two signatures come from two clusters of lyase conformations, each with a very different class of conformations of the bound NAD cofactor.

We found that the structural fold and the conformation of the bound NAD cofactor are the two major determinants of the formation of the clusters of the NAD binding pockets (Fig. 4a). It can be seen in Fig. 4b–j that there are two general conformations of the NAD coenzyme. The NAD coenzymes labeled C (Fig. 4b, c, f, g, h, and j) have a closed conformation, while the coenzymes labeled X (Fig. 4d, e, and i) have an extended conformation. This indicates that the binding pocket may take multiple conformations yet bind the same substrate in the same general structure. For example, the two structurally distinct signature pockets shown in Fig. 4f, g are derived from proteins that have the same biological function and SCOP fold. All of these proteins bind to the same NAD conformation.

We have further evaluated the effectiveness of the NAD binding site basis set by determining its accuracy in correctly classifying enzymes as either NAD-binding or non-NAD-binding. We constructed a test data set of 576 surface pockets from the CASTp database [60] independent of the training set of 457 NAD binding proteins. These 576 surface pockets were selected by taking the top 3 largest pockets in volume from 142 randomly chosen proteins and 50 proteins that have NAD bound in the PDB structure, with the further constraint that they were not in our training data set. We then structurally aligned all 576 pockets in our test data set against each of the nine NAD signature pockets in the resulting basis set. The testing pocket was assigned to be an NAD binding pocket if it structurally aligned to one of the nine NAD signature pockets, with its distance under a predefined threshold. Otherwise it

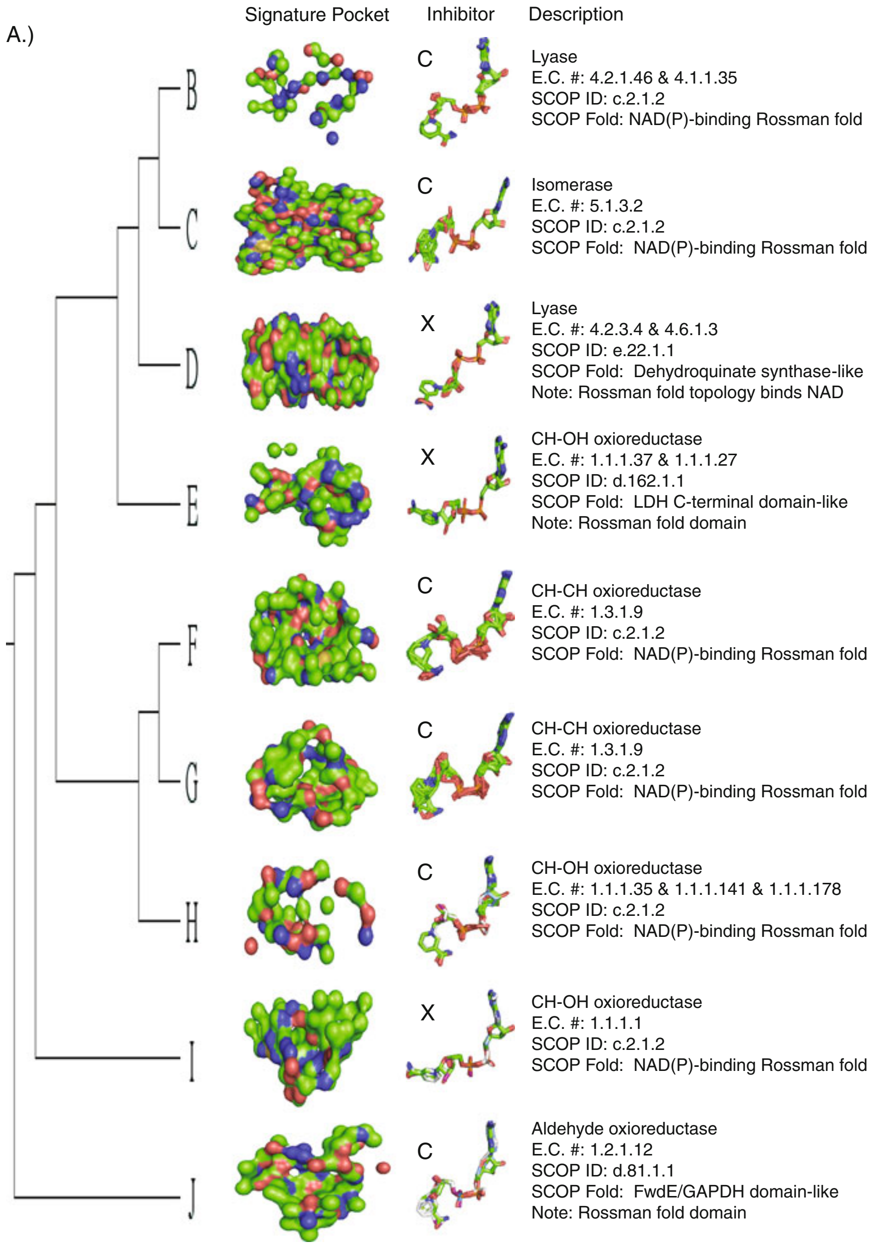
A.)
Signature Pocket
Inhibitor
Description
B
C
Lyase
E.C. #: 4.2.1.46 & 4.1.1.35
SCOP ID: c.2.1.2
SCOP Fold: NAD(P)-binding Rossman fold
C
C
Isomerase
E.C. #: 5.1.3.2
SCOP ID: c.2.1.2
SCOP Fold: NAD(P)-binding Rossman fold
D
X
Lyase
E.C. #: 4.2.3.4 & 4.6.1.3
SCOP ID: e.22.1.1
SCOP Fold: Dehydroquinate synthase-like
Note: Rossman fold topology binds NAD
E
X
CH-OH oxioreductase
E.C. #: 1.1.1.37 & 1.1.1.27
SCOP ID: d.162.1.1
SCOP Fold: LDH C-terminal domain-like
Note: Rossman fold domain
F
C
CH-CH oxioreductase
E.C. #: 1.3.1.9
SCOP ID: c.2.1.2
SCOP Fold: NAD(P)-binding Rossman fold
G
C
CH-CH oxioreductase
E.C. #: 1.3.1.9
SCOP ID: c.2.1.2
SCOP Fold: NAD(P)-binding Rossman fold
H
C
CH-OH oxioreductase
E.C. #: 1.1.1.35 & 1.1.1.141 & 1.1.1.178
SCOP ID: c.2.1.2
SCOP Fold: NAD(P)-binding Rossman fold
I
X
CH-OH oxioreductase
E.C. #: 1.1.1.1
SCOP ID: c.2.1.2
SCOP Fold: NAD(P)-binding Rossman fold
J
C
Aldehyde oxioreductase
E.C. #: 1.2.1.12
SCOP ID: d.81.1.1
SCOP Fold: FwdE/GAPDH domain-like
Note: Rossman fold domain

Fig. 4 The topology of the hierarchical tree and signature pockets of the NAD binding pockets. **a** The resulting hierarchical tree topology. **b–j** The resulting signature pockets of the NAD binding proteins, along with the superimposed NAD molecules that were bound in the pockets of the member proteins of the respective clusters. The NAD coenzymes have two distinct conformations. Those in an extended conformation are marked with an X and those in a compact conformation are marked with a C

was classified as non-NAD binding. The results show that the basis set of 9 signature pockets can classify the correct NAD binding pocket with sensitivity and specificity of 0.91 and 0.89, respectively. We performed further testing to determine whether a single representative NAD binding pocket, as opposed to a basis set, is sufficient for identifying NAD-binding enzymes. We chose a pocket representative pocket from one of the 9 clusters that were used to construct the 9 signature pockets. Here, a testing pockets was classified as NAD-binding if its structural similarity to the single representative pocket was above the same pre-defined threshold used in the basis set study. We repeat this exercise nine times, each time using a different representative from a different cluster. We found that the results deteriorated significantly, with an average sensitivity and specificity of only 0.36 and 0.23, respectively. This study strongly indicates that the construction of a basis set of signatures as a structural template provides significant improvement for a set of proteins binding the same co-factor but with diverse evolutionary origin. Further details of the NAD-binding protein study can be found in [66], along with an in-depth study of the metalloendopeptidase, including the construction of its signatures and basis set, as well as their utility in function prediction.

Conclusion

In this chapter, we have discussed methods that provide solutions to the problem of aligning protein global structures as well as aligning protein local surface pockets. Both methods disregard the ordering of residues in the protein primary sequences. For global alignment of protein structures, such a method can be used to address the challenging problem of identifying proteins that are topologically permuted but are spatially similar. The approach of fragment assembly based on the formulation of a relaxed integer programming problem and an algorithm based on scheduling split-interval graphs works well, and is characterized by a guaranteed approximation ratio. In a scaled up study, we showed that this method enables in discovery of circularly permuted proteins, including several previously unrecognized protein pairs. It also uncovered a case of two proteins related by higher order permutations.

We also described a method for order-independent alignment of local spatial surfaces that is based on bi-partite graph matching. By assessing surface similarity for a group of protein structures of the same function, this method can be used to automatically construct signatures and basis set of binding surfaces characteristic of a specific biological function. We showed that such signatures can reveal useful mechanistic insight on enzyme function, and can correlate well with substrate binding specificity.

In this chapter, we neglected an important issue in our discussion of comparing protein local surfaces for inferring biochemical functions, namely, how to detect evolutionary signals and how to employ such information for protein function prediction. Instead of going into details, we first point readers to the general approach of constructing continuous time Markovian models to study protein evolution [70, 71]. In addition, a Bayesian Monte Carlo method that can separate selection pressure due to biological function from selection pressure due to the constraints of protein folding stability and folding dynamics can be found in [57] and in [72]. The Bayesian Monte Carlo approach can be used to construct customized scoring matrices that are specific to a particular class of proteins of the same function. Details of how such method works and how it can be used to accurately predict enzyme functions from structure with good sensitivity and specificity for 100 enzyme families can be found in a recent review [72] and original publications [8, 57]. The task of computing surface pockets and voids using alpha shape is discussed in a recent review [73].

Acknowledgements This work was supported by NIH grants GM079804, GM081682, GM086145, NSF grants DBI-0646035 and DMS-0800257, ONR grant N00014-09-1-0028.

References

1. Binkowski, A., Joachimiak, A., Liang, J. Protein surface analysis for function annotation in high-throughput structural genomics pipeline. Protein Sci. **14**: 2972–2981 (2005).
2. Pazos, F., Sternberg, M.J.E. Automated prediction of protein function and detection of functional sites from structure. PNAS **101**:14, 14754–14759 (2004).
3. Andrade, M.A., Brown, N.P., Leroy, C., Hoersch, S., de Daruvar, A., Reich, C., Franchini, A., Tamames, J., Valencia, A., Ouzounis, C., Sander, C. Automated genome sequence anlysis and annotation. Bioinformatics **15**: 391–412 (1999).
4. Jensen, L.J., Gupta, R., Blom, N., Devos, D., Tamames, J., Kesmir, C., Nielsen, H., Staerfeldt, H.H., Rapacki, K., Workman, C., Andersen, C.A.F., Knudsen, S., Krogh, A., Valencia, A., Brunak, S. Prediction of human protein function from post-translational modifications and localization features. J. Mol. Biol. **319**: 1257–1265 (2002).
5. Pal, D., Eisenberg, D. Inference of protein function from protein structure. Structure **13**: 121–130 (2005).
6. Laskowski, R.A., Watson, J.D., Thornton, J.M. ProFunc: a server for predicting protein function from 3D structure. Nucleic Acids Res. **33**: W89–93 (2005).
7. Deng, M., Zhang, K., Mehta, S., Chen, T., Sun, F. Prediction of protein function using protein-protein interaction data. J. Comput. Biol. **10**(6): 947–960 (2003).
8. Tseng, Y.Y., Dundas, J., Liang, J. Predicting protein function and binding profile via matching of local evolutionary and geometric surface patterns. J. Mol. Biol. **387**(2): 451–464 (2009).
9. Shah, I., Hunterm, L. Predicting enzyme function from sequence: a systematic appraisal. ISMB **5**: 276–283 (1997).
10. Altschul, S.F., Warren, G., Miller, W., Myers, E.W., Lipman, D.J. Basic local alignment search tool. J. Mol. Biol. **215**: 403–410 (1990).
11. Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. **25**(17): 3389–3402 (1997).
12. Karplus, K., Barret, C., Hughey, R. Hidden Markov Models for detecting remote protein homologues. Bioinformatics **14**: 846–856 (1998).
13. Hulo, N., Sigrist, C.J.A., Le Saux, V. Recent improvements to the PROSITE database. Nucleic Acids Res. **32**: D134–D137 (2004).
14. Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M.C., Estreicher, A., Gasteiger, E., Martin, M.J., Michoud, K., O'Donovan, C., Phan, I., Pilbout, S., Schneider, M. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res. **31**: 365–370 (2003).

15. Weidong, T., Skolnick, J. How well is enzyme function conserved as a function of pairwise sequence identity. J. Mol. Biol. **333**: 863–882 (2003).
16. Rost, B. Twilight zone of protein sequence alignments. Protein Eng. **12**: 85–94 (1999).
17. Hegyi, H., Gerstein, M. The relationship between protein structure and function: a comprehensive survey with application to the yeast genome. J. Mol. Biol. **288**: 147–164 (1999).
18. Murzin, A.G., Brenner, S.E., Hubbard, T., Chothia, C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J. Mol. Biol. **247**: 536–540 (1995).
19. Orengo, C.A., Michie, A.D., Jones, D.T., Swindells, M.B., Thornton, J.M. CATH: a hierarchical classification of protein domain structures. Structure **5**: 1093–1108 (1997).
20. Holm, L., Sander, C. Protein structure comparison by alignment of distance matrices. J. Mol. Biol. **233**: 123–138 (1993).
21. Shindyalov, I.N., Bourne, P.E. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. Protein Eng. **11**(9): 739–747 (1998).
22. Hasegawa, H., Holm, L. Advances and pitfalls of protein structural alignment. Curr. Opin. Struct. Biol. **19**: 341–348 (2009).
23. Dayhoff, M.O., Schwartz, R.M., Orcutt, B.C. A model of evolutionary change in proteins. Atlas Protein Seq. Struct. **5**(3): 345–352 (1978).
24. Henikoff, S., Henikoff, J.G. Amino acid substitution matrices from protein blocks. PNAS **89**(22): 10915–10919 (1992).
25. Teichert, F., Bastolla, U., Porto, M. SABERTOOTH: protein structure comparison based on vectorial structure representation. BMC Bioinformatics **8**: 425 (2007).
26. Lindqvist, Y., Schneider, G. Circular permutations of natural protein sequences: structural evidence. Curr. Opin. Struct. Biol. **7**: 422–427 (1997).
27. Dundas, J., Binkowski, T.A., DasGupta, B., Liang, J. Topology independent protein structural alignment. BMC Bioinformatics **8**(388) doi:10.1186/1471-2105-8-388 (2007).
28. Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., Bourne, P.E. The protein data bank. Nucleic Acids Res. **28**: 235–242 (2000).
29. Umeyama, S. Least-squares estimation of transformation parameters between two point patterns. IEEE Trans. Pattern Anal. Mach. Intell. **13**(4): 376–380 (1991).
30. Veeramalai, M., Gilbert, D. A novel method for comparing topological models of protein structures enhanced with ligand information. Bioinformatics **24**(23): 2698–2705 (2008).
31. Aghili, S.A., Agrawal, D., El Abbadi, A. PADS: protein structure alignment using directional shape signatures. In DASFFA (2004).
32. Szustakowski, J.D., Weng, Z. Protein structure alignment using a genetic algorithm. Proteins: Struct. Funct. Genet. **38**: 428–440 (2000).
33. Standley, D.M., Toh, H., Nakamura, H. Detecting local structural similarity in proteins by maximizing number of equivalent residues. Proteins: Struct. Funct. Genet. **57**: 381–391 (2004).
34. Roach, J., Sharma, S., Kapustina, M., Cater Jr., C.W. Structure alignment via delaunay tetrahedralization. Proteins: Struct. Funct. Genet. **60**: 66–81 (2005).
35. Teyra, J., Paszkowski-Rogacz, M., Anders, G., Pisabarro, M.T. SCOWLP classification: structural comparison and analysis of protein binding regions. BMC Bioinformatics doi:10.1186/1471-2105-9-9 (2008).
36. Gold, N.D., Jackson, R.M. Fold independent structural comparisons of protein-ligand binding sites for exploring functional relationships. J. Mol. Biol. **355**: 1112–1124 (2006).
37. Zhu, J., Weng, Z. A novel protein structure alignment algorithm. Proteins: Struct. Funct. Bioinform. **58**: 618–627 (2005).
38. Bar-Yehuda, R., Halldorsson, M.M., Naor, J., Shacknai, H., Shapira, I. Scheduling split intervals. 14th ACM-SIAM Symposium on Discrete Algorithms, Baltimore, MD, pp. 732–741 (2002).

39. Hobohm, U., Sander, C. Enlarged representative set of protein structures. Protein Sci. **33**: 522 (1994).
40. Dutta, S., Akey, I.V., Dingwall, C., Hartman, K.L., Laue, T., Nolte, R.T., Head, J.F., Akey, C.W. The crystal structure of nucleoplasmin-core implication for histone binding and nucleosome assembly. Mol. Cell **8**: 841–853 (2001).
41. Woo, E.J., Marshall, J., Bauly, J., Chen, J.G., Venis, M., Napier, R.M., Pickersgill, R.W. Crystal structure of the auxin-binding protein 1 in complex with auxin. EMBO J. **21**: 2877–2885 (2002).
42. Uliel, S., Fliess, A., Amir, A., Unger, R. A simple algorithm for detecting circular permutations in proteins. Bioinformatics **15**(11): 930–936 (1999).
43. Alexandrov, N.N., Fischer, D. Analysis of topological and nontopological structural similarities in the PDB: new examples with old structures. Proteins **25**: 354–365 (1996).
44. Dror, O., Benyamini, H., Nussinov, R., Wolfson, H.J. MASS: multiple structural alignment by secondary structures. Bioinformatics **19**: i95–i104 (2003).
45. Shih, E.S., Hwang, M.J. Alternative alignments from comparison of protein structures. Proteins **56**: 519–527 (2004).
46. Ilyin, V.A., Abyzov, A., Leslin, C.M. Structural alignment of proteins by a novel TOPOFIT method, as a superimposition of common volumes at a topomax point. Protein Sci. **13**: 1865–1874 (2004).
47. Tabtiang, R.K., Cezairliyan, B.O., Grant, R.A., Cochrane, J.C., Sauer, R.T. Consolidating critical binding determinants by noncyclic rearrangement of protein secondary structure. PNAS **7**: 2305–2309 (2004).
48. Warren, A.J., Bravo, J., Williams, R.L., Rabbitts, T.H. Structural basis for the heterodimeric interaction between the acute leukemia-associated transcription factors AML1 and CBFbeta. EMBO J. **19**: 3004–3015 (2000).
49. Meining, W., Eberhardt, S., Bacher, A., Ladenstein, R. The structure of the N-terminal domain of riboflavin synthase in complex with riboflavin at 2.6A resolution. J. Mol. Biol. **331**: 1053–1063 (2003).
50. Lichtarge, O., Bourne, H.R., Cohen, F.E. An evolutionary trace method defines binding surfaces common to protein families. J. Mol. Biol. **7**: 39–46 (1994).
51. Norel, R., Fischer, H., Wolfson, H., Nussinov, R. Molecular surface recognition by computer vision-based technique. Protein Eng. **7**(1): 39–46 (1994).
52. Fischer, D., Norel, R., Wolfson, H., Nussinov, R. Surface motifs by a computer vision-technique: searches, detection, and implications for protein-ligand recognition. Proteins **16**: 278–292 (1993).
53. Meng, E., Polacco, B., Babbitt, P. Superfamily active site templates. Proteins **55**: 962–967 (2004).
54. Orengo, C., Todd, A., Thornton, J. From protein structure to function. Curr. Opin. Struct. Biol. **9**: 374–382 (1999).
55. Binkowski, A., Adamian, L., Liang, J. Inferring functional relationship of proteins from local sequence and spatial surface patterns. J Mol Biol. **332**: 505–526 (2003).
56. Jeffery, C. Molecular mechanisms for multi-tasking: recent crystal structures of moon-lighting proteins. Curr. Opin. Struct. Biol. **14**: 663–668 (2004).
57. Tseng, Y.Y., Liang, J. Estimation of amino acid residue substitution rates at local spatial regions and application in protein function inference: a Bayesian Monte Carlo approach. Mol. Biol. Evol. **23**: 421–436 (2006).
58. Liang, J., Edelsbrunner, H., Woodward, C. Anatomy of protein pockets and cavities: measurement of binding site geometry and implications for ligand design. Protein Sci. **7**: 1884–1897 (1998).
59. Edelsbrunner, H., Facello, M., Liang, J. On the definition and the construction of pockets in macromolecules. Disc Appl. Math. **88**: 83–102 (1998).
60. Dundas, J., Ouyang, Z., Tseng, J., Binkowski, A., Turpaz, Y., Liang, J. CASTp: computed atlas of surface topography of proteins with structural and topographical mapping of functionally annotated residues. Nucleic Acids Res. **34**: W116–W118 (2006).

61. Lee, S., Li, B., La, D., Fang, Y., Ramani, K., Rustamov, R., Kihara, D. Fast protein tertiary structure retrieval based on global surface shape similarity. Proteins **72**: 1259–1273 (2008).
62. Binkowski, T.A., Joachimiak, A. Protein functional surfaces: global shape matching and local spatial alignments of ligand binding sites. BMC Struct. Biol. **8**: 45 (2008).
63. Bandyopadhyay, D., Huan, J., Liu, J., Prins, J., Snoeyink, J., Wang, W., Tropsha, A. Functional Neighbors: Inferring relationships between non-homologous protein families using family-specific packing motifs. Proc. IEEE Int. Conf. Bioinform. Biomed. **14**(5): 1137–1143 (2008).
64. Mol, M., Kavraki, L.E. LabelHash: A flexible and extensible method for matching structural motifs. Automated Function Prediction Meetings, Toronto, Canada (2008).
65. Chen, L., Wu, L.Y., Wang, R., Wang, Y., Zhang, S., Zhang, X.S. Comparison of protein structures by multi-objective optimization. Genome Inform. **16**(2): 114–124 (2005).
66. Dundas, J. Adamian, L. Liang, J. Structural signatures of enzyme binding pockets from order-independent surface alignment: a study of metalloendopeptidase and nad binding proteins. J. Mol. Biol. **406**(5): 713–729 (2011 Mar).
67. Corment, T.H., Leiserson, C.E., Rivest, R.L., Stein, C. *Introduction to algorithms*, 2nd edn. Cambridge, MA: MIT Press (2001).
68. Kuhn, H.W. The hungarian method for the assignment problem. Nav. Res. Logist. Q. **2**: 83–97 (1995).
69. Bellman, R. On a routing problem. Q. Apply Math. **16**(1): 87–90 (1958).
70. Yang, Z., Nielsen, R., Hasegawa, M. Models of amino acid substitution and applications to mitochondrial protein structures. Mol. Biol. Evol. **15**: 1600–1611 (1998).
71. Huelsenbeck, J.B., Ronquist, R., Nielsen, R., Bollback, J. Bayesian inference of phylogeny and its impact on evolutionary biology. Science **294**: 2310–2314 (2001).
72. Liang, J., Tseng, Y.Y., Dundas J., Binkowski, A., Joachimiak, A., Ouyang, Z., Adamian, L. Chapter 4: predicting and characterizing protein functions through matching geometric and evolutionary patterns of binding surfaces. Adv. Protein Chem. **75**: 107–141 (2008).
73. Liang, J., Kachalo, S., Li, X., Ouyang, Z., Tseng, Y.Y., Zhang, J. Geometric structures of proteins for understanding folding, discriminating natives and predicting biochemical functions. *The World is a Jigsaw*. van de Weygaert R. (ed.). Springer (2009).