16 Algorithmic Methodologies for Discovery of Non-sequential Protein Structure Similarities

BHASKAR DASGUPTA

Department of Computer Science
University of Illinois at Chicago
Chicago, IL 60607
Email: bdasgup@uic.edu

JOE DUNDAS

Department of Bioengineering
University of Illinois at Chicago
Chicago, IL 60607
Email: jdunda1@gmail.com

JIE LIANG

Department of Bioengineering
University of Illinois at Chicago
Chicago, IL 60607
Email: jliang@uic.edu

16.1 INTRODUCTION

An increasing number of protein structures are becoming available that have either no known function or knowledge of the proteins functional mechanism is incomplete. Using experimental methods alone to explore these proteins in order to determine their functional mechanism is unfeasible. For this reason, much research has been put into computational methods for predicting the function of proteins [5, 31, 44, 34, 14, 53]. One such computational method is functional inference by homology where annotations from a protein with known function are transferred onto another protein based on sequence or structural similarities.

Protein sequence comparisons have been used as a straightforward method for functional inheritance. If two proteins have a high level of sequence identity, frequently the two proteins have the same or related biological function. This observation has been used as a basis for transferring annotations from a protein that is well characterized to a protein with unknown function when the two proteins have high sequence similarity [45, 4, 3]. It is often the case that only the protein residues that are near the functional region of the protein are under evolutionary pressure for conservation. Therefore, the global sequence similarity may be relatively low while local regions within the two sequences maintain a higher level of sequence similarity. In this case, probabilistic models such as profiles have been constructed using only the local regions of high sequence similarity [3, 32, 28]. Sequence comparison methods have the advantage that there are large numbers of sequences deposited into sequence databases such as SWISS-PROT [11] which provides adequate information for constructing probabilistic models. However, a relatively high level of sequence similarity is needed in order to accurately transfer protein function. In fact, problems begin to arise when the sequence identity between a pair of proteins is less than 60% [57].

Because proteins often maintain structural similarities even when sequence identity falls as low as 30% [6], making protein structure more strongly correlated with protein function than protein sequence [24]. For this reason, many researchers have begun comparing the three-dimensional structure of proteins in an attempt to uncover more distant evolutionary relationships among proteins. The SCOP [40] and CATH [43] databases have organized protein structures hierarchically into different classes and folds based on their overall similarity in topology and fold. Classification of protein structures relies heavily on the reliable protein structure comparison methods. Common structural comparison methods include Dali [27] and CE [47]. However, structural alignment methods cannot guarantee optimal results and do not have an interpretability comparable to sequence alignment methods.

Several challenges arise when trying to compare protein structures. First, when searching for global structural similarity, similar to sequence alignment methods, one can search for global similarity or similarity within local surface regions of interest. Unlike sequence alignment scoring methods which are heavily based on models of protein evolution [13, 25], scoring systems for structural alignment must take into account both the three-dimensional positional deviations between the aligned residues or atoms, as well as other biologically important shared characteristics. Defining a robust quantitative measure of similarity is challenging. The difficulty is illustrated by the variety of structural alignment scoring methods that have been proposed [23]. Second, many alignment methods assume the ordering of the residues follows that of the primary sequence [47, 51]. This sequence order dependence can lead to problems when comparing local surface regions which often have residues and atoms from different locations on primary sequence fold together to form functional regions in three-dimensional space. On the global backbone level, the existence of permuted proteins, such as the circular permutation [37, 17] also poses significant problems for sequence order dependent alignment methods. Third, proteins may undergo small side chain structural fluctuations or larger backbone fluctuations *in vivo* which are not represented in a single static snapshot of a crystallized structure in the Protein

Data Bank [9]. Many structural alignment methods assume rigid bodies and cannot take structural changes into account.

In this chapter, we will discuss several issues of structural alignment and then discuss methods we have implemented for sequence order independent structural alignment at the global level and at the local surface level. We illustrate the utility of our methods by showing how our sequence order independent global structural alignment method detects circular permuted proteins. We then show how our local surface sequence order independent structural alignment method can be used to construct a basis set of signature pockets of binding surfaces for a specific biological function. The signature pocket represents structurally conserved surface regions. A set of signature pockets can then be used to represent a functional family of proteins for protein function prediction.

16.2 STRUCTURAL ALIGNMENT

Comparing the structure of two proteins is an important problem [23] that may detect evolutionary relationships between proteins even when sequence identity between two proteins is relatively low. A widely used method for measuring structural similarity is the root mean squared distance (RMSD) between the equivalent atoms or residues of two proteins. If the equivalence relationship is known, a rotation matrix R and a translation vector T that when applied to one of the protein structures will minimize the RMSD can be found by solving the minimization problem:

$$min \sum_{i=1}^{N_B} \sum_{j=1}^{N_A} |T + RB_i - A_j|^2, \tag{16.1}$$

where N_A is the number of points in structure A and N_B is the number of points in structure B. If $N_A = N_B$ then the least-squares estimation of the parameters R and T in Eq. 16.1 can be found using singular value decomposition.

The equivalence relationship is usually not known *a priori* when aligning to protein structures. In this case, the structural alignment method is a problem of minimizing RMSD while maximizing the number of aligned points. Heuristics must be used to solve this multi-objective optimization problem.

A number of heuristic methods have been developed [56, 1, 49, 48, 52, 22, 59] that can be divided into two main categories. *Global* methods are used to detect similarities between the overall fold of two proteins and *local* alignment methods are used to detect similarities within local regions of interest within the two proteins. Most current methods are restricted to only finding structural similarities where the order of the structural elements within the alignment follow the order of the elements within the primary sequence. Sequence order independent methods ignore the sequential ordering of the atoms or residues in primary sequence. These methods are better suited for finding more complex global similarities and can also be employed for finding all atoms local comparisons. We have implemented both sequence order independent methods for both global and local structural alignments.

16.3 GLOBAL SEQUENCE ORDER INDEPENDENT STRUCTURAL ALIGNMENT

Looking for similarities between the overall fold can elucidate evolutionary or functional relationships between two proteins. However, most of the current methods for structural comparison are sequence order dependent and are restricted to comparing similar topologies between the two backbones. It has been discovered that throughout evolution, a genetic event can rearrange the topology of the backbone. One such example is the circular permutation. Conceptually, a circular permutation can be thought of as a ligation of the N- and C-termini of a protein and cleavage somewhere else on the protein backbone. It has been observed that circular permutations often maintain a similar three-dimensional spatial arrangement of secondary structures. In addition to circular permutations, research has shown that more complex topological rearrangements are possible [37]. Detection of these permuted proteins will be valuable for studies in homology modeling, protein folding, and for protein design.

16.3.0.1 Sequence Order Independent Global Structural Alignment We have developed a sequence order independent structural alignment algorithm for detecting structural similarities between two protein that have undergone topological rearrangement of their backbone structures [17]. Our method is based on fragment assembly where the two proteins to be aligned are first exhaustively fragmented. Each fragment $\lambda^{A}_{i,k}$ from protein structure S_A is pair-wise superimposed onto each fragment $\lambda^{B}_{j,k}$ from protein structure S_B. The result is 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 fragments. The variable $k \in \{5, 6, 7\}$ is the length of the fragment. Each fragment pair is assigned a similarity score,

$$\sigma(\chi_{i,j,k}) = \alpha[C - s(\chi_{i,j,k}) \cdot \frac{cRMSD}{k^2}] + SCS, \tag{16.2}$$

where $cRMSD$ is the measured RMSD value after optimal superposition, 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 the fragments, and SCS is a BLOSSUM-like measure of similarity in sequence of the matched fragments [25]. Details of the scoring method can be found in [17].

The goal of the structural alignment is to find a consistent set of fragment pairs $\Delta = \{\chi_{i_1,j_1,k_1}, \chi_{i_2,j_2,k_2}, ..., \chi_{i_t,j_t,k_t}\}$ that minimizes the overall 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 the scheduling of split-interval graphs [8] and is itself based on a fractional version of the local-ratio approach.

To begin, a conflict graph $G = (V, E)$ is created, where a vertex is defined for each aligned fragment pair. Two vertices are connected by an edge if any of the fragments $(\lambda^{A}_{i,k}, \lambda^{B}_{i‘,k‘})$ or $(\lambda^{B}_{j,k}, \lambda^{B}_{j‘,k‘})$ from the fragment pair is not disjoint, that is, if both fragments from the same protein share one or more residues. For each vertex representing aligned fragment pairs, we assign three indicator variables $x_\chi, y_{\chi_{\lambda_A}}$, and $x_\chi, y_{\chi_{\lambda_B}} \in \{0, 1\}$ and a closed neighborhood $Nbr[\chi]$.x_χ indicates whether the fragment pair should be used ($x_\chi = 1$) or not ($x_\chi = 0$) in the final alignment. $x_\chi, y_{\chi_{\lambda_A}}$ and $x_\chi, y_{\chi_{\lambda_B}}$ 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 connect to χ by an edge.

The sequence order independent structural alignment algorithm can be described as follows. To begin, 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
   $$\sum_{\chi \in \Delta} \sigma(\chi) \cdot x_\chi \tag{16.3}$$
   subject to
   $$\sum_{a_t \in \lambda^A} y_{\chi_{\lambda_A}} \leq 1 \quad \forall a_t \in S_A \tag{16.4}$$
   $$\sum_{b_t \in \lambda^B} y_{\chi_{\lambda_B}} \leq 1 \quad \forall b_t \in S_B \tag{16.5}$$
   $$y_{\chi_{\lambda_A}} - x_\chi \leq 1 \quad \forall \chi \in \Delta \tag{16.6}$$
   $$y_{\chi_{\lambda_B}} - x_\chi \leq 1 \quad \forall \chi \in \Delta \tag{16.7}$$
   $$x_\chi, y_{\chi_{\lambda_A}}, y_{\chi_{\lambda_B}} \leq 1 \quad \forall \chi \in \Delta \tag{16.8}$$

2. For every vertex $\chi \in V_\Delta$ of G_Δ, compute its *local conflict number* $\alpha_\chi = \sum_{\chi‘ \in Nbr_\Delta[\chi]} x_{\chi‘}$. Let χ_{min} be the vertex with the *minimum* local conflict number. Define a new similarity function σ_{new} from σ as follows:
   $$\sigma_{new}(\chi) = \begin{cases} \sigma(\chi), & \text{if } \chi \notin Nbr_\Delta[\chi_{min}] \\ \sigma(\chi) - \sigma(\chi_{min}) & \text{otherwise} \end{cases}$$

3. Create $\Delta_{new} \subseteq \Delta$ by removing from Δ every substructure pair χ such that $\sigma_{new} \leq 0$. Push each removed substructure on to a stack in arbitrary order.

4. If $\Delta_{new} \neq 0$ then repeat from step 1, setting $\Delta = \Delta_{new}$ and $\sigma = \sigma_{new}$. 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 exists 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 guarantees that each substructure from a substructure pair is in the same spatial arrangement in the global alignment.

16.3.0.2 Detecting Permuted Proteins This algorithm was implemented in a large scale study to search for permuted proteins in the Protein Data Bank [9]. A subset of 3,336 protein structures taken from the PDBSELECT90 data set [26] are structurally aligned in a pair-wise fashion. From the subset of 3,336 proteins, we aligned two proteins if they met the following conditions (see [17] for details):

1. The difference in their lengths was no more than 75 residues.
2. The two proteins shared approximately the same secondary structure content.

Within the approximately 200,000 structural alignments performed, we found many known circular permutations and three novel circular permutations, as well as a more complex pair of non-cyclic permuted proteins. Here we describe the details of the circular permutation 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 We found a novel circular permutation between the nucleoplasmin-core protein in *Xenopu laevis* (PDB ID `1k5j`, chain E) [19] and the auxin binding protein in maize (PDB ID `1lrh`, chain A, residues 37-127) [58]. The structural alignment between `1k5jE` (Fig. 16.1, top) and `1lrhA` (Fig. 16.1, 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 the p-value calculation can be found in [17]. The short loop connecting two antiparallel strans in nucleoplasmin-core protein (in circle, top of Fig. 16.1b) becomes disconnected in auxin binding protein 1 (in circle, bottom of Fig. 16.1b), and the N- and C- termini of the nucleoplasmin-core protein (in square, top of Fig. 16.1b) are connected in auxin binding protein 1 (square, bottom of Fig. 16.1b). For details of other circular permutations we found, including permutations between microphage migration inhibition factor and the C-terminal domain of arginine repressor, please see [17].

Complex Protein Permutations Because of their relevance in understanding the functional and folding mechanism of proteins, circular permutations have received much attention [37, 55]. However, the possibility of more complex backbone rearrangements were experimentally verified by artificially rearranging the topology of the ARC repressor and were found to be thermodynamically stable[50]. Very little is known about this class of permuted proteins, and the detection of non-cyclic permutations is a challenging task [2, 15, 46, 29].

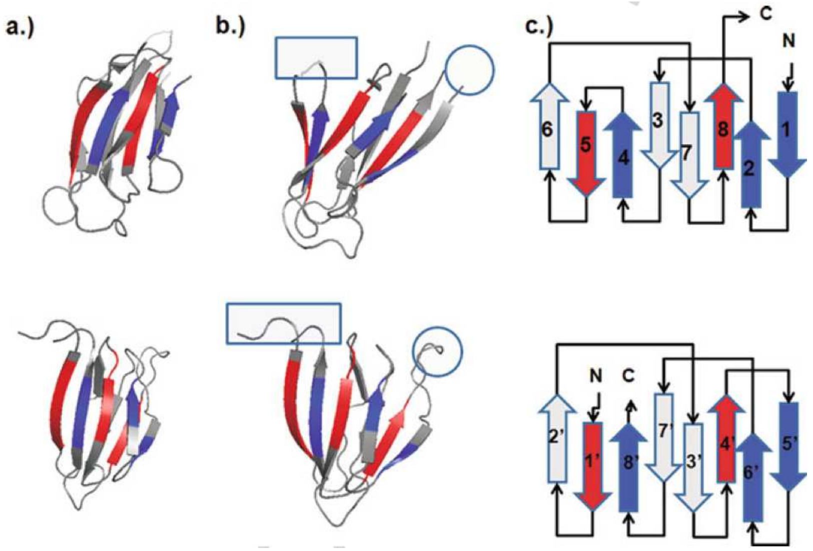

Fig. 16.1 A newly discovered circular permutation between nucleoplasmin-core (`1k5j`, chain E, *top panel*), and a fragment of auxin binding protein 1 (residues 37-127) (`1lrh`, chain A, *bottom panel*). a. These two proteins align well with a RMSD value of 1.36 Åover 68 residues, with a significant 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 int eh PDB structure file. This figure is modified from [17].

Our database search uncovered a naturally occurring non-cyclic permutation between chain F of AML1/Core binding factor (AML1/CBF, PDB ID `1e50`, Fig. 16.2a, top) and chain A of riboflavin synthase (PDB ID `1pkv`, Fig. 16.2a, bottom). The two structures align well with an RMSD of 1.23 Åat 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. 16.2b) can be transformed into that of riboflavin synthase (Fig. 16.2f) 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. 16.2c). Connect the C-terminal end of strand 4 to the original N-terminal (Fig. 16.2d). Connect the C-terminal end of strand 5 to the N-terminal end of helix 2 (Fig. 16.2e). 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. 16.2f).

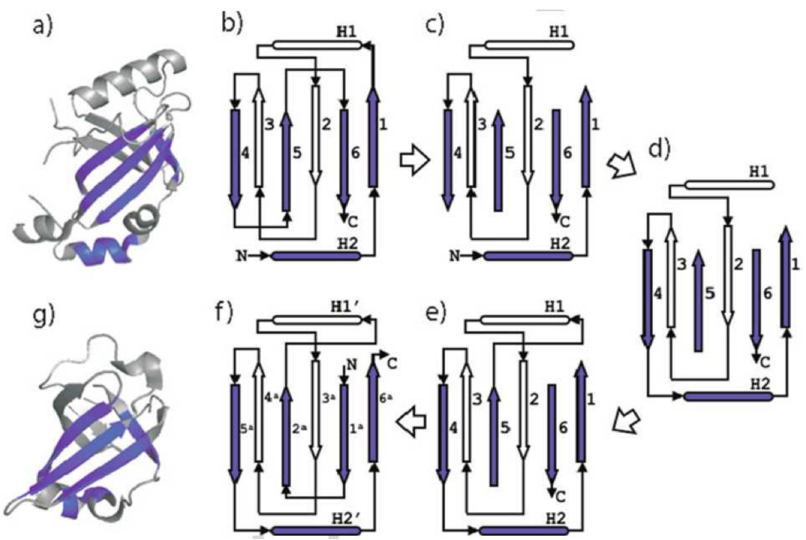

Fig. 16.2 A non-cyclic permutation discovered between AML1/Core Binding Factor (AML1/CBF PDB ID 1e50, Chain F, *top*) and riboflavin synthase (PDB ID 1pkv, chain A, *bottom*). **a** These two proteins structurally align with an RMSD of 1.23 Åover 42 residues and has a significant p-value of 2.8×10^{-4} after Bonferroni correction. The residues that were assigned equivalences from the structural alignment are colored blue. **b** These proteins are structurally 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 is modified from [17].

16.4 LOCAL SEQUENCE ORDER INDEPENDENT STRUCTURAL ALIGNMENT

The comparison of the global backbone can lead to discovery of distant evolutionary relationships between proteins. However, when attempting to detect similar functions or functional mechanisms between two proteins, global backbone similarity is not a robust indicator [36, 41, 20]. It can be assumed that the physiochemical properties of the local region where function takes place (i.e. substrate binding) is under more evolutionary pressure to be conserved. This assumption has been recently backed up by several studies [53, 38, 42, 21, 30, 54].

A typical protein contains many concave surface regions, commonly referred to as *surface pockets*. However, only a few of the surface pockets supply a unique physiochemical environment that is conducive for the protein to carry out its function. The protein must maintain this surface pocket throughout evolution in order to conserve 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 [21, 35, 10, 7, 39]

The inherent flexibility of the protein structure makes the problem of structural comparison of protein surface pockets challenging. A protein is not a static structure as represented by a Protein Data Bank [9] entry. The whole protein as well as the local functional surface may undergo various degrees of structural fluctuations. The use of a single surface pocket structure as a representative template for a specific protein function can lead to many false negatives. This is due to the inability of a single representative to capture the full functional characteristics across all conformations of a protein.

We have addressed this problem by developing an algorithm that can identify the atoms of a surface pocket that are structurally preserved across a family of protein structures that have similar function. Using a sequence order independent local surface alignment method to pair-wise align the functional pockets across a family of protein structure, we automatically find the structurally conserved atoms and measure their fluctuations. We call these structurally conserved atoms the *signature pocket*. More than one signature pocket may result for single functional class. In this case, our method can automatically create a *basis set* of signature pockets for that functional family. These signature pockets can then be used as representatives for scanning a structure database for functional inference by structural similarity.

16.4.1 Bi-partite Graph Matching Algorithm for Local Surface Comparison

We have modified and implemented a sequence order independent local structural alignment algorithm based on the maximum weight bi-partite graph matching formulation developed by Chen et.al. [12].

As mentioned earlier, the structural alignment problem bares down to a problem of finding an equivalence relation between residues of a reference protein S_R and a query protein S_Q that when applied will optimize the superposition of the two structures. The formulation here does this in an iterative two step process. First, an optimal set of equivalent atoms are determined under the current superposition using a bi-partite graph representation. Second, the new equivalence relation is used to determine a new optimal superposition. The two steps are then repeated until a stopping condition is met.

The equivalence relationship is found between two the atoms of the functional pocket surfaces by representing the atoms the atoms of S_R and S_Q as nodes in a graph. This graph is bi-partite, meaning that edges only exist between atoms of S_R and atoms of S_Q. A directed edge is drawn between two nodes if a similarity threshold is met. In our implementation, the measure of similarity takes into account both spatial distances and the chemical property similarities between the two corresponding atoms.

Each edge $e_{i,j}$ connecting node i and node j is assigned a weight $w(i,j)$ equal to the similarity score between the two corresponding atoms (see [] for details). The optimal equivalence relationship given the current superposition is a subset of the edges within this bi-partite graph that have maximum combined weight, where at most one edge can be selected per atom making this a maximum-weight bi-partite graph matching problem. The solution to this problem can be found using the Hungarian algorithm [33].

The Hungarian method is as follows. Initially, an overall score $F_{all} = 0$ is set. Additionally, an artificial source node s and an artificial destination node d are added to the bipartite graph. A directed edge es, i with zero weight is added for each of the atom nodes i from S_R and similarly, directed edges ej, d with zero weight are drawn from each of the atoms nodes of S_Q. 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 [] algorithm.
2. Assign a new weight $w'(i,j)$ to each edge that does not originate from the source node s as follows,

$$w'(i,j) = w(i,j) + [F(i) - F(j)]. \tag{16.9}$$

3. Update F_{all} as $F'_{all} = F_{all} - F_d$.
4. Reverse the direction of the edges along the shortest path from s to d.
5. If $F_{all} > F_d$ and a path exists between s and d then go back to step 1.

The iterative process of the Hungarian algorithm stops 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}. At the end of the process, the graph will consist of a set of directed edges that have been reversed (they now point from nodes of S_Q to nodes from S_R. These reversed edges represent the new equivalence relationships between the atoms of S_Q and the atoms of S_R.

The equivalence relationship found by the bi-partite matching algorithm can now be used to superimpose the two proteins using the singular value decomposition. After superpositioning the new equivalent atoms, a new bi-partite graph is created and the process is iterated until the change in $RMSD$ upon superposition falls below a threshold.

16.4.2 A Basis Set of Binding Surface Signature Pockets

Being able to compare structural similarities between to protein surface regions can provide insight into shared biological functions. As mentioned earlier, when dealing with local surface regions one has to be careful when choosing a functional representative pocket due to the inherent flexibility of the binding surfaces. We have developed a method that automatically generates a set of functional pocket templates, called *signature pockets* of local surface regions that can be used as a representative a functional surface for structural comparison. These signature pockets contain broad structural information and have discriminating ability.

A signature pocket is derived from sequence order independent structural alignments of precomputed surface pockets. Our signature pocket 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 hierarchically organized according to their structural similarity.

The input of our signature pocket algorithm is a set of functional pockets from the CASTp database [18]. All vs all pair-wise sequence order independent local surface alignment is performed on the input functional surface pockets. A distance is calculated based on the RMSD and the chemistry of the paired atoms of the structural alignment (see [16] for details). The resulting distance matrix is used by an agglomerative clustering method. The signature of the functional pocket is then derived using a recursive process following the hierarchical tree.

The recursive process begins by finding the two closest siblings (pockets S_A and S_B), and combining them into a single structure S_{AB}. During the recursive process, S_A or S_B may themselves already be a combination of several structures. When combining two structures, we follow the criteria 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 as the average of the two underlying atom coordinates.

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.

A count of the number of times an atom at the position i was present in the underlying set of pockets (N) is recorded during each step in the recursive process. A *preservation ratio* $\rho(i)$ is calculated for each atom of the signature pocket by dividing N by the total number of constituent pockets. In addition, the mean distance of the coordinates of the aligned atoms to their geometric center is recorded as the *location variation* υ. At the end of each step, the new structure S_{AB} replaces the two structures S_Aa nd S_B in the hierarchical tree and the process is repeated on the updated hierarchical tree.

The recursive process can be stopped at any point during its traversal of the hierarchical tree by selecting a ρ threshold. Depending on the choice of the ρ threshold, a single or multiple signature pockets can be created. Figure 16.3a shows a low ρ threshold which results in a set of 3 signature pockets. As the threshold is raised, fewer signature pockets are created (Fig. 16.3b). A single signature pocket representing all surface pockets in the data set can be generated by raising the threshold even further (Fig. 16.3). The set of signature pockets from different clusters in the hierarchical tree form a *basis set* that represents an ensemble of differently sampled conformations of the surface pockets in the Protein Data Bank. The basis set of signature pockets can be used to accurately classify and predict enzymes function.

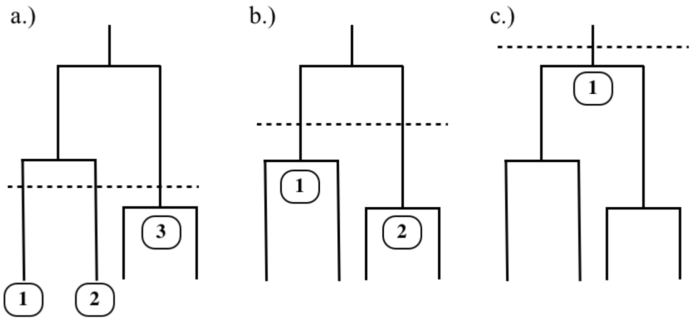

Fig. 16.3 Different basis sets of signature pockets can be produced at different levels of structural similarity by raising or lowering the similarity threshold (*vertical dashed line*). **a** A low threshold will produce more signature pockets. **b** As the threshold is raised, fewer signature pockets 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.

16.4.2.1 Signature Pockets of NAD Binding Proteins To illustrate how signature pockets and the basis set help to identify structural elements that are important for binding and to show their accuracy in functional inference, we discuss a study performed on the nicotinamide adenine dinucleotide (NAD) bindign proteins. NAD plays essential roles in metabolisms where it acts as a coenzyme in redox reactions, including glycolysis and the citric acid cycle.

We obtained a set of 457 NAD binding proteins of diverse fold and diverse evolutionary origin. We extracted the NAD binding surfaces from the CASTp database of protein pockets [18]. We obtained the hierarchical tree using the results of our sequence order independent surface alignments. The resulting 9 signature pockets of the NAD binding pocket form a basis set, shown in Figure16.4.

The signature pockets of NAD contain biological information. The signature pocket show in Figure 16.4j is based on a cluster of NAD binding proteins that act on the aldehyde group of donors, the signature pockets in (Fig. 16.4f and g) are for oxioreductases that act on the CH-CH group of donors, and the signature pockets of Figure 16.4e, h, and i are for clusters of alcohol oxioreductases that act on the CH-OH group of donors. The NAD-binding lyase family is represented in two signature pockets. The first represents lyases that cleave both C-O and P-O (Fig. 16.4d) and the second containing lyases that cleave both C-O and CC bonds (Fig. 16.4b). These two signature pockets from two clusters of lyase conformations have a different class of conformations of the bound NAD cofactor (extended and compact).

In addition to the structural fold, the signature pockets are also determined by the conformation of the bound NAD cofactor (Fig. 16.4a). It can be seen in Fig. 16.4b-j that there are two general conformations of the NAD coenzyme. The coenzymes labeled C (Fig. 16.4b,c,f,g,h, and j) have a closed conformation, while the conenzymes labeled X (Fig. 16.4d,e and i) have an extended conformation. This indicates that the binding pocket may take multiple conformations yet bind the same substrate in the

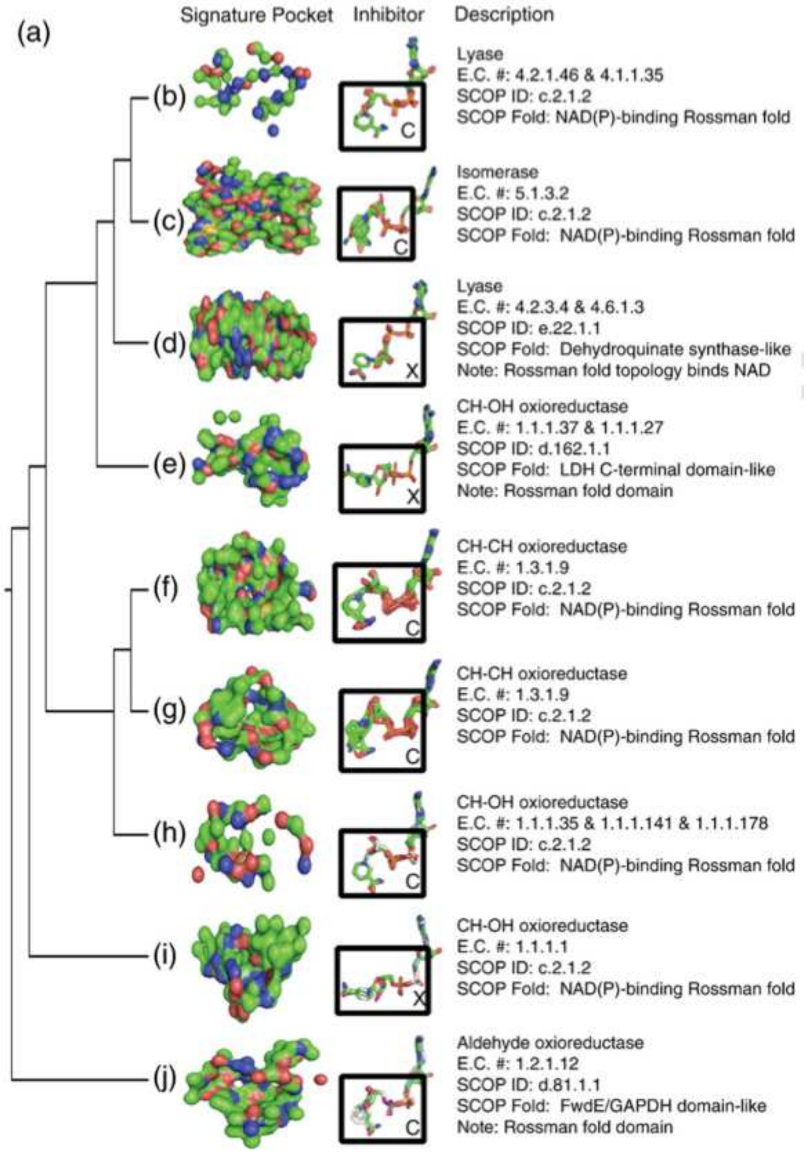
(a)
Signature Pocket
Inhibitor
Description
(b)
C
Lyase
E.C. #: 4.2.1.46 & 4.1.1.35
SCOP ID: c.2.1.2
SCOP Fold: NAD(P)-binding Rossman fold
(c)
C
Isomerase
E.C. #: 5.1.3.2
SCOP ID: c.2.1.2
SCOP Fold: NAD(P)-binding Rossman fold
(d)
X
Lyase
E.C. #: 4.2.3.4 & 4.6.1.3
SCOP ID: e.22.1.1
SCOP Fold: Dehydroquinate 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. 16.4 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.

same general structure. For example, the two structurally distinct signature pockets shown in Fig. 16.4f and 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 further evaluated the effectiveness of the NAD basis set by determining its accuracy at correctly classifying enzymes as either NAD-binding or non-NAD-binding. We constructed a testing data set of 576 surface pockets from the CASTp database [18]. This data set is independent of the 457 NAD binding proteins we used to create the signature pockets. We collected the 576 surface pockets by selecting the top 3 largest pockets by volume from 142 randomly chosen proteins and 50 proteins that have NAD bound in the PDB structure. We then structurally aligned each of the signature pockets against each of the 576 testing pockets. The testing pocket was assigned to be an NAD binding pocket if it structurally aligned to one of the nine NAD signature pockets with a distance under a predefined threshold. Otherwise it 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 single pocket representative from one of the nine clusters at random and attempted to classify our testing data set by structural alignment. We used the same predefined threshold used in the basis set study. This was repeated 9 times using a representative from each of the 9 clusters. We found that the results deteriorated significantly with an average sensitivity and specificity of 0.36 and 0.23, respectively. This strongly indicates that the construction of a basis set of signature pockets to be used as a structural template provides significant improvement for functional inference of a set of evolutionarily diverse proteins.

16.5 CONCLUSION

We have discussed methods that provide solutions to the problems that arise during functional inference by structural similarity at both the global level and at the local surface level. Both of our methods disregard the ordering of residues in the proteins primary sequence, making them sequence order independent. The global method can be used to address the challenging problem of detecting structural similarities even after topological rearrangements of the proteins backbone. The fragment assembly approach based on the formulation of a relaxed integer programming problem and an algorithm based on scheduling split-interval graphs is guaranteed by an approximation ratio. We showed that this method is capable of discovering circularly permuted proteins and other more complex topological rearrangements.

We also described a method for sequence order independent alignment of local surfaces on proteins. This method is based on a bi-partite graph matching problem. We further show that the surface alignments can be used to automatically construct a basis set of signature pockets representing structurally preserved atoms across a family of proteins with similar biological function.

Acknowledgments

DasGupta and Liang was partially supported by NSF grant DBI-1062328.

REFERENCES

1. AS Aghili, D Agrawal, and A El Abbadi. PADS: protein structure alignment using directional shape signatures. *In DASFFA*, 2004.
2. NN Alexandrov and D Fischer. Analysis of topological and nontopological structural similarities in the PDB: new examples from old structures. *Proteins*, 25:354–365, 1996.
3. SF Altschul, TL Madden, AA Schaffer, J Zhang, Z Zhang, W Miller, and Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, 25:3389–3402, 1997.
4. SF Altschul, G Warren, W Miller, EW Myers, and Lipman DJ. Basic local alignment search tool. *J. Mol. Biolo.*, 215:403–410, 1990.
5. MA Andrade, Brown NP, Leroy C, S Hoersch, A de Daruvar, C Reich, J Franchini, A an dTamames, A Valencia, C Ouzounis, and SAnder C. Automated genome sequence analysis and annotation. *Bioinformatics*, 15:391–412, 1999.
6. Rost B. Twilight zone of protein sequence alignments. *Protein Eng.*, 12:85–94, 1999.
7. D Bandyopadhyay, J Huan, J Liu, J Prins, J Snoeyink, W Wang, and A Tropsha. Functional neighbors: Inferring relationships between non-homologous protein families using family specific packing motifs. *Proc. IEEE int. Conf. Bioinform. Biomed.*, 2008.
8. R Bar-Yehuda, MM Halldorsson, J Naor, H Shacknai, and I Shapira. Scheduling split intervals. *14th ACM-SIAM Symposium on Discrete Algorithms*, pages 732–741, 2002.
9. HM Berman, J Westbrook, Z Feng, G Gilliland, TN Bhat, H Weissig, IN Shindyalov, and PE Bourne. The protein data bank. *Nucleic Acids Res.*, 28:235–242, 2000.
10. TA Binkowski and A Joachimiak. Protein functional surfaces: global shape matching and local spatial alignments of ligand binding sites. *BMC Struct. Biol.*, 8:45, 2008.
11. B Boeckmann, A Bairoch, R Apweiler, Blatter MC, A Estreicher, Gasteiger E, MJ Martin, K Michoud, C O'Donovan, I Phan, S Pilbout, and M Schneider. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.*, 31:365–370, 2003.

12. L Chen, LY Wu, R Wang, Y Wang, S Zhang, and XS Zhang. Comparison of protein structures by multi-objective optimization. *Genome Inform.*, 16(2):114–124, 2005.

13. MO Dayhoff, RM Schwartz, and BC Orcutt. A model of evolutionary change in proteins. *Atlas Protein Seq. Struct.*, 5(3):345–352, 1978.

14. M Deng, K Zhang, S Mehta, T Chen, and F Sun. Prediction of protein function using protein-protein interaction data. *J. Comput. Biol.*, 10(6):947–960, 2009.

15. O Dror, H Benyamini, R Nussinov, and HJ Wolfson. MASS: multiple structural alignment by secondary structures. *Bioinformatics*, 19:i95–i104, 2003.

16. J Dundas, L Adamian, and J Liang. Signatures and basis sets of enzyme binding surfaces by sequence order independent surface alignment. *J. Mol. Biol*, doi:10.1016/j.jmb.2010.12.005, 2010.

17. J Dundas, TA Binkowski, B DasGupta, and J Liang. Topology independent protein structural alignment. *BMC Bioinformatics*, 8:388 doi:10.1186/1471–2105–8–388, 2008.

18. J Dundas, Z Ouyang, J Tseng, TA Binkowski, Y Turpaz, and J Liang. CASTp: computed atlast of surface topography of proteins with structural and topographical mapping of functionally annotated residues. *Nucleic Acids Res.*, 34:W116–W118.

19. S Dutta, IV Akey, C Dingwall, KL Hartman, T Laue, Nolte RT, Head JF, and CW Akey. The crystal structure of nucleoplasmin-core implications for histone binding and neucleosome assemply. *Mol. Cell*, 8:841–853, 2001.

20. D Fischer, R Norel, H Wolfson, and R Nussinov. Surface motifs by a computer vision technique: searches, detection, and implications for protein-ligand recognition. *Proteins*, 16:278–292, 1993.

21. Inferring functional relationship of proteins from local sequence and spatial surface patterns. J. mol. biol. *332*, 332:505–526, 2003.

22. ND Gold and RM Jackson. Fold independent structural comparisons of protein-ligand binding sites for exploring functional relationships. *J. Mol. Biol.*, 355:1112–1124, 2006.

23. H Hasegawa and L Holm. Advances and pitfalls of protein structural alignment. *Curr. Opin. Struct. Biol.*, 19:341–348, 2009.

24. H Hegyi and M Gerstein. The relationship between protein structure and function: a comprehensive survey with application to the yeast genome. *J. Mol. Biol.*, 288:147–164, 1999.

25. S Henikoff and JG Henikoff. Amino acid substitution matrices from protein blocks. *PNAS*, 89(22):10915–10919, 1992.

26. Hobohm U, Sander C: Enlarged representative set of protein structures. *Protein Science* 1994, **3**:522.

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

28. N Hulo, CJA Sigrist, and V Le Saux. Recent improvements to the PROSITE database. *Nucleic Acids Res.*, 32:D134–D137, 2004.

29. VA Ilyin, A Abyzov, and CM Leslin. 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.

30. C Jeffery. Molecular mechanisms for multi-tasking; recent crystal structures of moon-lighting proteins. *Curr. Opin. Struct. Biol.*, 14:663–668, 2004.

31. LJ Jensen, R Gupta, N Blom, D DEvos, J Tamames, C Kesmir, H Nielsen, HH Staerfeldt, K Rapacki, C Workman, Andersen CAF, S Knudsen, A Krogh, A VAlencia, and S Brunak. Prediction of human protein function from post-translational modifications and localization features. *J. Mol. Biol.*, 319:1257–1265, 2002.

32. K Karplus, C Barret, and R Hughey. Hidden Markov Models for detecting remote protein homologues. *Bioinformatics*, 14:846–856, 1998.

33. HW Kuh. The hungarian method for the assignment problem. *Nav. Res. Logist. Q.*, 2:83–97, 1995.

34. RA Laskowski, JD Watson, and JM Thornton. ProFunc: a server for predicting protein function from 3D structure. *Nucleic Acids Res.*, 33:W89=93, 2005.

35. S Lee, B Li, D La, Y Fang, K Ramani, R Rustamov, and D Kihara. Fast protein tertiary structure retrieval based on global surface shape similarity. *Proteins*, 72:1259–1273, 2008.

36. O Lichtarge, HR Bourne, and Cohen FE. An evolutionary trace method defines binding surfaces common to protein families. *J. Mol. Biol.*, 7:39–46, 1994.

37. Y Lindqvist and Schneider G. Circular permutations of natural protein sequences: structural evidence. *Curr. Opin. Struct. Biol.*, 7:422–477, 1997.

38. E Meng, B Polacco, and P Babbitt. Superfamily active site templates. *Proteins*, 55:962–967, 2004.

39. M Moll and LE Kavraki. A flexible and extensible method for matching structural motifs. *Nat. Proc.*, 2008.

40. AG Murzin, SE Brenner, T Hubbard, and C Chothia. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, 247:536–540, 1995.

41. R Norel, H Fischer, H Wolfson, and R Nussinov. Molecular surface recognition by computer vision-based technique. *Protein Eng.*, 7(1):39–46, 1994.

42. C Orengo, A Todd, and J Thornton. From protein structure to function. *Curr. Opin. Struct. Biol.*, 9:374–382, 1999.

43. CA Orengo, AD michie, DT JOnes, MB Swindells, and JM Thornton. CATH: a hierarchical classification of protein domains structures. *Structure*, 5:1093–1108, 1997.

44. D Pal and D Eisenberg. Inference of protein function from protein structure. *Structure*, 13:121–130, 2005.

45. I Shah and L Hunterm. Predicting enzyme function from sequence: a systematic appraisal. *ISMB*, 5:276–283, 1997.

46. ES Shih and MJ Hwang. Alternative alignments from comparison of protein structures. *Proteins*, 56:519–527, 2004.

47. IN Shindyalov and PE Bourne. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng.*, 11(9):739–747, 1998.

48. DM Standley, H Toh, and H Nakamura. Detecting local structural similarity in proteins by maximizing the number of equivalent residues. *Proteins: Struct. Funct. Genet.*, 57:381–391, 2004.

49. JD Szustakowski and Z Weng. Protein structure alignment using a genetic algorithm. *Proteins: Struct. Func. Genet.*, 38:428–440, 2000.

50. RK Tabtiang, BO Cezairliyan, RA Grant, JC Chochrane, and RT Sauer. Consolidating critical binding determinants by noncyclic rearrangement of protein secondary structure. *PNAS*, 7:2305–2309, 2004.

51. F Teichert, U Bastolla, and M Porto. SABERTOOTH: protein structure comparison basedon vectorial structure representation. *BMC Bioinformatics*, 8:425, 2007.

52. J Teyra, M Paszkowski-Rogacz, G Anders, and MT Pisabarro. SCOWLP classification: structural comparison and analysis of protein bindign regions. *BMC Bioinformatics*, doi:10.1186/1471-2105-9-9, 2008.

53. YY Tseng, J Dundas, and J Liang. Predicting protein function and binding profile via matching of local evolutionary and geometric surface patterns. *J. Mol. Biol.*, 387(2):451–464, 2009.

54. YY Tseng and J Liang. 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.

55. S Uliel, A Fliess, A Amir, and R Unger. A simple algorithm for detecting circular permutations in proteins. *Bioinformatics*, 15(11):930–936, 1999.

56. M Veeramalai and D Gilbert. A novel method for comparing topological models of protein structures enhanced with ligand information. *Bioinformatics*, 24(23):2698–2705, 2008.

57. T. Weidong and J Skolnick. How well is enzyme function conserved as a function of pairwise sequence identity. *J. Mol. Biol.*, 333:863–882, 2003.

58. EJ Woo, J Marshall, J Bauly, JG Chen, M Venis, RM Napier, and RW Pickersgill. Crystal structure of the auxin-binding protein 1 in complex with auxin. *EMBO J.*, 21:2877–2885, 2001.

59. J Zhu and Z WEng. A novel protein structure alignment algorithm. *Proteins: Struct. Funct. Bioinform.*, 58:618–627, 2005.