Profile Combinatorics for Fragment Selection in Comparative Protein Structure Modeling

Deacon Sweeney  
*Wright State University - Main Campus*

Travis E. Doom  
*Wright State University - Main Campus*, travis.doom@wright.edu

Michael L. Raymer  
*Wright State University - Main Campus*, michael.raymer@wright.edu

Follow this and additional works at: https://corescholar.libraries.wright.edu/knoesis

Part of the Bioinformatics Commons, Communication Technology and New Media Commons, Databases and Information Systems Commons, OS and Networks Commons, and the Science and Technology Studies Commons

Repository Citation
https://corescholar.libraries.wright.edu/knoesis/532

This Conference Proceeding is brought to you for free and open access by the The Ohio Center of Excellence in Knowledge-Enabled Computing (Kno.e.sis) at CORE Scholar. It has been accepted for inclusion in Kno.e.sis Publications by an authorized administrator of CORE Scholar. For more information, please contact library-corescholar@wright.edu.
Profile Combinatorics for Fragment Selection in Comparative Protein Structure Modeling

Deacon J. Sweeney¹, Gerald M. Alter², Michael L. Raymer¹, Travis E. Doom¹
¹Department of Computer Science and Engineering
²Department of Biochemistry
Wright State University,
Dayton, OH 45435-0001
{sweeney.2, gerald.alter, michael.raymer, travis.doom} @wright.edu

Abstract

The sequencing of the human genome was a great stride toward modeling our cellular complexes, massive systems whose key players are proteins and DNA. A major bottleneck limiting the modeling process is structure and function annotation for the new genes. Contemporary protein structure prediction algorithms represent the sequence of every protein of known structure with a profile to which the profile of a protein sequence of unknown structure is compared for recognition. We propose a novel approach to increase the scope and resolution of protein structure profiles. Our technique locates equivalent regions among the members of a structurally similar fold family, and clusters these regions linkers by structural similarity. Equivalent substructures can then be swapped on the common regions to generate an array of profiles which represent hypothetical structures to supplement profiles of known structures. Strategies for a specific implementation of the strategy are discussed, including application to multiple template comparative modeling.

1 Introduction

Automated prediction offers a means for researchers and scientists to replace unnecessarily expensive and time consuming empirical methods for protein structure and function determination with rapid algorithmic predictions. Using an atomic level structure prediction for a protein, empirical protein structure data can be transformed into a comparative model of the protein [1-5]. Fragment-based principles have previously been applied toward high resolution comparative modeling of antibodies. A similar approach is here generalized to enhance the specificity of recognition of classes of structurally similar proteins, or fold families. The fragment-based approach to fold family recognition (fold recognition), lends itself to other applications, including enhanced prediction and classification of protein functions and a unique protein function optimization technique.

2 Background

2.1 Conservation and analogy

Functionally important amino acids that are structurally equivalent among the members of a fold family are sometimes required by natural selection to contain similar amino acid types. Evolutionary maintenance of the amino acid type at a specific position is called conservation of the property that the amino acid types have in common, for example a functionally important negatively charged amino acid in an enzyme’s active site might conserve an aspartic acid in all members of the ancestral family that perform the function. Analogy is used to refer to non-ancestrally related members of the same fold family achieving a similar function by convergent evolution. Thus by conservation and analogy analyses, a priori fold family and function classification [6] and prediction of protein-protein interfaces [7] are feasible for cases of great diversity even before a structural mechanism for the attribute has been proposed.

2.2 Modeling using sequence and structure

Modeling applications that combine sequence with structure information [8-10], such as 3D-pssm in fold
2.3 Canonical structures in antibody modeling

For the special case of antibodies, canonical structures have previously been defined that largely account for the structural variations that occur among the proteins of this type. The term canonical structure makes reference to subsets of structurally similar loops in the complementarity determining region (CDR) of a set of antibody protein structures [17-21]. Among the repertoire of antibody proteins in a human body, the CDR region is a set of six loops that achieve extremely high sequence diversity to recognize any non-self molecules. The high diversity and the vast number of sequences that have been determined make the antibody family an interesting target for well-informed comparative modeling studies. An approach has previously been delineated for selection of canonical structures for fragment-based comparative modeling of antibodies [22]. The approach involves the collection of the sequences of individual canonical structures into profiles which are stored for later use in recognizing new loops. To the best of our knowledge no attempt has ever been made to generalize this fragment cluster profiling approach to cover the canonical structures of other fold families. In our lab we are automating this procedure for all fold families. Our goal is to construct profiles for all combinations of canonical structures available for each fold family. We believe that the added diversity of hypothetical combinations may drastically increase the resolution of recognition in fold and function prediction for many cases, as is observed in antibody modeling when canonical structures are predicted independently from each other [22].

2.4 Key residues in canonical structures

Canonical structure modeling rests on the premise that the classification of any protein fold family’s sub-fold fragments into structurally similar groups automatically collects information on key residues [23-26] that are the primary determinants of the local topology [22,27]. In effect, with the classification of substructures one isolates the major part of the reason for their realization – the sequence pattern of the specific fragment. Later we capture other determinants of cluster selection, in what are called complementary misconservation patterns (CMP).

When a target protein’s profile is being compared to a template protein’s profile, alignment of similar profile positions drives recognition scores up. One should expect alignment of similar key residues to significantly influence the scores when comparing a target protein’s profile to a whole set of template profiles for a fold family. Recombination of canonical structure profiles into unique hypothetical profiles will serve to expand the scope of fold family profiles. Recombination will be done to facilitate selection of a highest scoring combination of canonical structures to assign as a structure prediction to a target protein’s profile. We are creating a situation where the choice of canonical structure combinations is not limited to those that have been realized in known structures. The different key residues that characterize canonical structures are expected to make a significant difference in recognition scores for selection of the best combination.

2.5 Complementary misconservation patterns

Recombination of structurally-linked fragment profiles can also support inference on coordinated events that increases the specificity of recognition for any given profile. The common regions of a fold family should not be expected to conserve precisely the same characteristics per combination of fragment profiles. Rather, common regions tend to vary slightly depending on their interactions with other fragments, primarily in levels of hydrophobic conservation in certain ‘core’ positions, as demonstrated in this paper. Hydrophobic residues ‘hide’ from water with other hydrophobic residues, and they tend to be conserved in positions where this molecular ‘glue’ increases the stability of the protein. For example, bulky canonical structures that come into contact with one or more other fragments should be expected to drive up the hydrophobic conservation at certain positions of the contacted fragments, because this ‘glue’ tends to stabilize their interaction and thus the stability of the fold. Profile permutations can be endowed with slightly different hydrophobic levels in the common positions, depending on hydrophobic conservation correlations observed between the fragment clusters present among the empirically derived structures. Non-local sequence conservation rules, called CMPs, can be learned in this way to support or weaken confidence in recognition of canonical structures for loop regions.

The combinatoric approach to profile construction, and
inclusion of CMPs, can be implemented in the context of multiple fold recognition algorithms for comparative modeling of the results. The direction our lab is taking is discussed, and includes an introduction to CMPare, an algorithm that uses the combinatoric approach to perform fragment-based comparative modeling. A protein function optimization application that might be implemented with CMPare is also discussed. Given a set of fold family members with a common function, the approach uses linear regression to select the optimal fragments from a set of fold family members, similar to the way regression is used in quantitative structure activity relationships (QSAR), to enhance a drug’s activity.

3 Methods

3.1 The structure dataset

An overview of the algorithm is presented in Figure 1. The 23 protein structures used in this study, taken from the CATH database [28] are 1asyA1, 1awlA2, 1b8aA1, 1bcpB, 1bcpD, 1bcpF, 1bovA, 1cuk1, 1esfA1, 1jmcA1, 1jmcA2, 1kawA, 1lt5D, 1mjC, 1pfsA, 1pysB, 1tiiD, 2sns, 3ullA, 1otcA1, 1otcA2, 1otcA3, identified by CATH nomenclature. The selection is meant to broadly represent the diverse OB-fold family. Terminal sequences were often deleted to simplify visualization. A multiple sequence alignment was generated using ClustalW1.7 [29], and analyzed by an in-house utility program to detect sequence identity between all pairs of sequences. The program compares all aligned positions that are not both gaps, for every pair of sequences. 3ullA and 1cuk1 (CATH nomenclature) are 32 percent identical, indicating an evolutionary relationship, but the rest of the pairs each have less than 15 percent identity shared, which means that any possible ancestral relationship are too distant to be detected by conventional ‘homology’ methods. Several of the proteins are expected to be remote homologs (ancestrally related proteins of less than thirty percent identity), including the bcp family and the combined jmc and otc families [6] (among others), but we have manually investigated the structures and have found a great deal of diversity, even among these families. To generate a structure alignment, the core definition of OB-folds [30,31] was used to locate, by hand, core residues that are present in every member of the fold family, to be used in aligning the structures. Ten positions from the multiple sequence alignment were selected for alignment of the OB-folds to 1jmcA2, which was chosen as a reference structure because it contains many of the most commonly observed core positions.

3.2 Defining common core regions

The multiple sequence alignment and the structure alignment were analyzed simultaneously by CMPare to detect aligned amino acids that share less than 2.5 Å root mean square (RMS) difference among their alpha carbons. These are considered core regions because they reliably hold the main hydrophobic positions that constitute the core of the protein. The program reports any continuous core stretches and ignores insignificant gaps. The terminal residues of these common clusters are considered transition points that can be used to allow interchange of loops among the set of family members, to formulate new combinations for profiling.

3.3 Clustering variable secondary structure units

Next, the gaps between common core regions as well as the C- and N-termini of the protein fold are searched for secondary structure units that are more variable than can be resolved by the initial scan for fragments with less than 2.5 Å RMS difference among their alpha carbon atom coordinates. Classification of secondary structure units is important because equivalent secondary structure units from a fold family often do not exhibit a common structure. Clustering of variable secondary structure units consists of locating, from the set of analogous fragments, subsets of high secondary structure content that exhibit a common local structural motif. The terminal fragments of each secondary structure cluster (common or variable) should next be sub-clustered amongst themselves.
3.4 Clustering terminal fragments

After all of the secondary structure units have been defined for the fold family, their terminal fragments are clustered by structure similarity. The clustering is done by hand at this time, in the following manner: Using Dino [32] to visualize multiple structures, the members of a target fragment set are compared to each other, all-against-all. For each terminal fragment, if a similar fragment is found, then this new ‘hit’ is recorded as a member of the cluster and taken out of the searchable set. To maximize the specificity of the clusters, previous hits are returned to the searchable set if a significant new hit includes significant alignment of a larger number of positions.

3.5 Correlation analysis

Two types of correlation analyses are then performed on the fragment clusters. First an occurrence correlation is measured between each of the variable fragments of the canonical structures that are being classified. The occurrence correlation measures the degree to which specific fragments appear together in the same parent structure. High occurrence correlations point out fragments that tend to occur together. These dependencies should be expected to be highest when two fragments physically interact with each other or are otherwise constrained to favor particular cluster combinations over others. For example, if one loop is significantly larger in one protein than in another, then another adjacent loop might also be longer than usual, if the two loops pack against each other. This correlation is measured to reduce the number of combinations necessary for generation of profiles, as highly linked clusters should also appear together in the profiles. Continuing with the example, the long loop might not be expected to occur if the adjacent loop were short, so this sort of mixed alignment would be scored with lower confidence. Erroneous correlations will occur if correlations are calculated for close homologs, which are more likely to contain similar fragment clusters due to their close ancestry than to the fragments’ packing.

Another correlation analysis searches for CMPs, non-local effects that the presence of specific fragment clusters have on the conservation observed in structurally adjacent fragments. To explain the collection of CMPs, some positions of any given fragment might be expected to be more or less conserved hydrophobic, depending on the burial state induced for the positions by the side-chains of adjacent fragments. In the above example of a long loop, the fact that the two loops pack together implies that certain positions in the second, structurally adjacent loop can be expected to be conserved hydrophobic, if the selection of the long-looped canonical structure is correct in the first place.

3.6 Conservation visualization

To simplify analyses of differences in conservation in fragments adjacent to a variable fragment, an in-house program called ColCorr, written in Sybyl Programming Language (SPL) [33], was used to locate residues that are buried and conserved hydrophobic. Side-chain percent accessibility was calculated using Naccess [34], and side-chains 35% or less accessible to solvent were considered buried for this study. A position is considered conserved hydrophobic if 60% or more of the positions in a column of a Blast sequence alignment are types \{A, C, F, G, I, L, M, V, W, Y\}. Residues both conserved hydrophobic and buried are colored white, conserved hydrophobic and accessible are colored cyan, not conserved hydrophobic and buried are colored yellow, and not conserved hydrophobic and accessible are colored green. ColCorr writes a Dino output file for high quality rendering of the properties.

4 Results

Four common fragments were located for the OB-folds analyzed, corresponding to beta strands one through four from Murzin’s description of the OB-fold [30]. Two secondary structure units, beta strand five and the alpha helix (from Murzin’s description) appearing between strands three and four did not contain residues used in the structure alignment. These fragments were not observed to have fragments with alpha carbons constrained to less than 2.5 Å RMS deviation across the entire fold family, which means that the fold family needs sub-classification by this secondary structure unit. The four common fragments that were located are shown in Figure 2.

![Figure 2. The four common fragments of OB-folds with positions with less than 2.5 Å RMS across all members of the family. Beta strand One is shaded darker to assist visualization.](image)
Manual execution of the clustering method was performed on a limited set of OB-fold fragments to exemplify the results that could be expected from an automated method. Manual clustering focused on the helix between strands three and four since it contains a secondary structure unit that diverges among family members. Three classes were assigned in the region at the absolute N-terminus of the fragment, which corresponds to the C-terminal fragment of beta strand three. The three clusters (Figure 3) included 14 of the folds. The rest of the family members have unique fragments in this region, and are not shown for clarity.

![Image](image_url)

**Figure 3. Clustering of the C-terminal fragment of the third beta strand.** Three clusters are visible for the fragment, textured with dark, medium, and light patterns. The untextured region absent in Figure 1 is the rest of the linker between strands three and four.

Three clusters were found in the alpha helix region (Figure 4a), and 6 were unique, again not shown. Note that these three clusters are of different sizes from and include different family members than the three clusters in Figure 3, i.e. their occurrence correlation is not high. The six unique members are also mostly from different parents. The three clusters are individually visualized in Figures 4b, 4c, and 4d. The N-termini of these clusters take largely unclusterable conformations, and again do not appear to correlate to the clustering of the C-termini of beta strand three. The C-termini of the helical region’s clusters are largely well clustered (also in 4b, c, d), and their occurrences appear to correlate with the clustering of the N-termini of beta strand four. These fragments (helical C-termini and strand four N-termini) are patterned consistently in 4b and 4c, to emphasize their high occurrence correlation.

An N-terminal alpha helix was observed in several of the OB-folds. The presence of this secondary structure unit has a substantial effect on the conservation of core residues of the beta strands forming the barrel of the OB-fold (Figure 5). When present, this helix inserts two hydrophobic side-chains into the barrel, capping it like the helix between strands three and four caps the other end of the barrel. When the helix is not present, a higher degree of polarity is observed in the conservation at positions with which the hydrophobic side-chains of the helix, when present, does come into contact (Table 1). We consider this non-local conservation a CMP. A similar correlation is observed in the occurrence of a C-terminal helix (not shown), and is expected to occur to a lesser degree with the different helices located between strands three and four. Such phenomena can be expected to occur upon drastic modifications in core secondary structure units - when structural variations occur, CMPs can be extracted from structurally adjacent fragments.

![Image](image_url)

**Figure 4. Clustering variable secondary structure units.** (a) shows three major clusters of a variable alpha helix distinguished by a diagonal pattern or medium or dark shading; (b) isolates the structures in the diagonally patterned cluster from (a), Three terminal clusters were recognized for this cluster, one shaded medium on the left of the cluster, and two on the right, shaded medium and dark. (c) shows the medium shaded cluster from (a), now diagonally patterned, with four terminal clusters, one with medium shading on the right and three with light, medium, and dark shading on the right. (d) shows the vertically textured cluster from (a), now diagonally patterned.
<table>
<thead>
<tr>
<th></th>
<th>a-N</th>
<th>B11</th>
<th>B12</th>
<th>B13</th>
<th>B32</th>
<th>B44</th>
<th>B45</th>
<th>B51</th>
<th>B52</th>
</tr>
</thead>
<tbody>
<tr>
<td>1aw7A2</td>
<td></td>
<td>w</td>
<td>w</td>
<td>y</td>
<td>y</td>
<td>y</td>
<td>g</td>
<td>y</td>
<td>y</td>
</tr>
<tr>
<td>1bcpD</td>
<td>absent</td>
<td>b</td>
<td>g</td>
<td>y</td>
<td>w</td>
<td>w</td>
<td>-</td>
<td>w</td>
<td></td>
</tr>
<tr>
<td>1bcpF</td>
<td>absent</td>
<td>g</td>
<td>b</td>
<td>w</td>
<td>w</td>
<td>w</td>
<td>g</td>
<td>-</td>
<td>w</td>
</tr>
<tr>
<td>1bovA</td>
<td>absent</td>
<td>w</td>
<td>g</td>
<td>g</td>
<td>y</td>
<td>g</td>
<td>g</td>
<td>-</td>
<td>y</td>
</tr>
<tr>
<td>1cu1k1</td>
<td>absent</td>
<td>b</td>
<td>w</td>
<td>w</td>
<td>b</td>
<td>y</td>
<td>y</td>
<td>g</td>
<td>w</td>
</tr>
<tr>
<td>1esf</td>
<td>absent</td>
<td>w</td>
<td>g</td>
<td>g</td>
<td>y</td>
<td>g</td>
<td>g</td>
<td>-</td>
<td>y</td>
</tr>
<tr>
<td>1kaw1</td>
<td>absent</td>
<td>w</td>
<td>w</td>
<td>y</td>
<td>w</td>
<td>b</td>
<td>g</td>
<td>w</td>
<td></td>
</tr>
<tr>
<td>1lt5D</td>
<td>absent</td>
<td>g</td>
<td>w</td>
<td>w</td>
<td>g</td>
<td>-</td>
<td>-</td>
<td>y</td>
<td></td>
</tr>
<tr>
<td>1mjc</td>
<td>absent</td>
<td>-</td>
<td>w</td>
<td>w</td>
<td>b</td>
<td>-</td>
<td>-</td>
<td>y</td>
<td></td>
</tr>
<tr>
<td>1psfs</td>
<td>absent</td>
<td>b</td>
<td>w</td>
<td>w</td>
<td>b</td>
<td>w</td>
<td>g</td>
<td>b</td>
<td>w</td>
</tr>
<tr>
<td>1 pysB</td>
<td>absent</td>
<td>b</td>
<td>b</td>
<td>w</td>
<td>w</td>
<td>y</td>
<td>-</td>
<td>-</td>
<td>w</td>
</tr>
<tr>
<td>1tiiD</td>
<td>absent</td>
<td>g</td>
<td>b</td>
<td>w</td>
<td>b</td>
<td>w</td>
<td>-</td>
<td>-</td>
<td>g</td>
</tr>
<tr>
<td>2 bcpB2</td>
<td>absent</td>
<td>y</td>
<td>w</td>
<td>w</td>
<td>w</td>
<td>y</td>
<td>g</td>
<td>w</td>
<td></td>
</tr>
<tr>
<td>3ullA</td>
<td>absent</td>
<td>w</td>
<td>w</td>
<td>y</td>
<td>w</td>
<td>b</td>
<td>g</td>
<td>y</td>
<td></td>
</tr>
</tbody>
</table>

**Table 1. Conservation differences observed in fragments near the N-terminal helical region, depending on the presence/absence of the helix.** The letters represent colors as described in Section 3.6, where w stands for white and represents buried and conserved hydrophobic, b stands for blue and represents solvent accessible and conserved hydrophobic, y stands for yellow and represents buried and conserved polar, g stands for green and represents solvent accessible and conserved polar. The column titles refer to specific positions as shown in Figure 4.

### 5 Discussion

This paper demonstrates an algorithmic approach to a general classification scheme for sub-fold topological clustering. The purpose for this is two-fold. First, profiles can be generated to cover hypothetical protein folds that are not represented in the set of known structures. The approach should reliably generate realistic profiles, as their constituents are built from fragments that are actually realized in the set of known structures. The approach relies on fragments that do not interact with each other being replaceable if the replacement set also does not interact. It seems likely that this will be useful to fold recognition because it allows an algorithmic approach to fragment selection, specifying more intimately what profile features would best represent a sequence, rather than simply assigning gross similarity to a single template structure.

This study shows evidence that clustering occurs in unrelated proteins. This is probably due to local constraints that compel fragments to evolve into the observed clusters. Secondary structure units’ terminal fragments, and even some of the secondary structure units themselves may randomly and independently relocate about the prominent clusters and occasionally achieve novel possibilities as the proteins evolve. The fact that the evolution of loops is often a random walk implies that different combinations of the observed example clusters should be realized by any given member of the fold family.

The second purpose of this work is to demonstrate that, by sub-fold clustering, profiles can be unambiguously in a

**Figure 5.** Residues whose conservation changes depending on the presence or absence of the N-terminal helix. The structure is 1jmcA2, and the helix is colored red. Eight positions are highlighted. Each affected position is labeled with three characters. A B (for beta) is followed by the strand number (i.e. B1 = strand one), which is followed by the order of the specific position in that strand (i.e. B11 = the N-terminal most core position of strand one).
manner both local and non-local in sequence, to represent particular combinations of fragments with higher specificity in generated profiles. Recognition of a specific fragment cluster in one region of an alignment in addition to preferential recognition of a CMP-adjusted profile in one or more adjacent fragments should be given higher confidence than either of the events as might occur separately. In fact, the alternative profiles not being found in adjacent fragments might be a key to recognizing that an alternative explanation for the canonical structure exists.

Perhaps the most compelling attribute of the proposed approach is that it makes transparent the incorporation of multiple templates into a single comparative model for a protein. In fact, the algorithm should normally be expected to select fragments from multiple templates when modeling gaps in structures. In our implementation, the program can output alpha carbon coordinates for the multiple templates selected for modeling, so construction of a comparative model of any given profile can be performed by any program that is capable of regenerating a full protein backbone structure from alpha carbon coordinates.

6 Conclusion

We are writing code to perform alignments between combinatorially generated profiles and profiles that we will generate to represent target proteins of unknown structure. Our intended approach to alignment is in the style of Taylor’s MST algorithm, which uses local and pairwise scoring to find an alignment that best fills hydrophobic cores as they occur in proteins of known structure. The algorithm is most suitable for recognizing remote homology and analogy relationships, and is not specific to pure sequence comparisons. The MST pseudo-side-chain beta carbon approximation approach allows the alignment algorithm the flexibility required for use with our recombined fragments. Also, this approach will allow unrealistic combinations to be discarded if their alpha-carbon distances are judged to be too unrealistic [35]. We will optimize the method by aligning and scoring both MST- and 3D-pssm style to find the most reliable in the context of our work. Separation of alignment and scoring is prominent in threading algorithms, so a combined approach should not be disqualified without merit.

Another potential application of the current approach is toward function optimization in protein engineering. Specific fragment sets of a protein can be targeted with optimization strategies analogous to drug optimization methods. Given binding strengths or activities for this set of proteins, a quantitative structure activity relationship (QSAR) could be constructed, where analogous fragments are represented as values of variables that generate an activity constant. Another approach from drug design that might be implemented in CMPare is application of D-optimal design [36] toward construction of a minimal set of profiles that adequately represents any given fold family for fold recognition. Briefly, the algorithm asks – if the target is a member of this family, approximately what profile would best represent it? The actual recognition score for this fold family is then extracted from the target’s alignment to this optimized profile. The approach can thus be used with rapid scoring for increased resolution fold recognition that can cover all fold families with a feasible time complexity.

References


[33] SYBYL: SYBYL 6.7.1 Tripos Inc., 1699 South Hanley Rd., St. Louis, Missouri, 63144, USA.

