# The Discovery of Structurally Conserved Regions by Dihedral Angles of Protein Backbone Sequence

Hui-Yu Shen
Department of Computer and Communication Engineering,
Chienkuo Institute of Technology
wyshen@ckit.edu.tw

## Abstract

The structurally conserved regions are usually obtained by pairwise or multiple sequence alignment (MSA). The commonly aligned protein sequence segments are referred to as evolutionarily remained residues which constitute the structurally conserved regions. Although its direct insight into the relationship between protein sequence and structure, sequence alignment is restricted to be performed from left to right residue, i.e., from N' to C' terminus, or vice versa. But in reality, the tertiary structure alignments do not mention any aligned direction. In other words, the protein sequence may be aligned with another sequence which may be in a totally reversed direction. Hence, we present an algorithm to calculate the structurally conserved regions directly from the dihedral angles of protein backbone sequence instead of sequence alignment. We are convinced that the proposed algorithm behaves not only concrete and extensive but also a qualified measure index for tertiary structure alignment.

**Keywords**：Sequence Alignment, Structure Alignment, Structurally Conserved Regions.

## 1. Introduction

Sequence alignment [1, 2] is usually used to determine the similarity between query and subject protein sequences. As a result, after sequence alignment operation, the evolutionary relationship, i.e., phylogenetic tree, can be derived by calculating the mutation distance with substitution matrix such as BLOSUM or PAM. But it is still a Holy Grail to forecast the complete protein tertiary structure only based on sequence information. Hence, some predictive algorithms such as simulated annealing [3] and genetic algorithm [4, 5] are devoted to achieve a reasonable estimate for protein tertiary structure. Once an estimated 3D model is obtained, it is time to give a measure index to assay the predicted model. RMSD (Root Mean Square Deviation) is always adopted to play such a role by computing the Euclidean distance between target protein and estimated model.

## 2. Motivation

Whenever sequence alignment is performed, each alphabet is either shifted if mismatch occurs or in correspondence with another alphabet depending on the scoring matrix. In such a way, it should be observed that sequence alignment always runs from left to right direction. But in 3D space, it has no any meaning about left or right direction. For example, if we have two sequence segments such as CDEFGHI and IHGFEDC, there is at most only one matched alphabet after using sequence alignment. But in 3D world, the matched alphabets may reach to seven as shown in Fig. 1. Furthermore, even there exist two identical sequence segments which are distantly located by each other, the sequence alignment operation may still fail to align these two identical sequence segments. That is, some gaps should be inserted between them. Nevertheless, they may match well in 3D space as depicted in Fig. 2. Therefore, it is not sufficient to identify the structurally conserved regions when we only carry out sequence alignment. In other words, some structurally conserved regions will be lost if only sequence information is well known.

## 3. Algorithm

The dihedral angles of backbone conformation, ? and f, can be calculated directly from their corresponding PDB files if the coordinates are well known for each residue. It is worth noted

that the backbone conformation is totally determined by these two dihedral angles. In other words, once ? and f for each residue are calculated, we can uniquely identify an exact protein sequence. Here, we divide the algorithm into the following three steps and start from these two dihedral angles.

(1) Relation Graph

Remember that the planar conformation constituted by peptide bond between *Ca* atoms is almost rigid. Therefore, it is sufficient to describe the overall backbone structure only by the dihedral angles, *?* and $f$. Let the vector $v_i = [?_i, f_i]^T$ stand for dihedral angles of backbone structure for some *Ca* atom of residue $r_i$. The backbone structure can be determined by the set $\{v^s_i\}$ for each protein $p_s$ as explained above. Suppose we have two protein backbone structures with $m$ and $n$ residues respectively, $\{v^s_i\}$ and $\{v^t_j\}$ where $i=1\sim n$ and $j=1\sim m$, to be compared with each other. Whenever if $\|v^s_i - v^t_j\| = e$ for some predefined distance e, then an edge $e_k$ is allocated between these two vectors, $v^s_i$ and $v^t_j$. Each edge between $v^s_i$ and $v^t_j$ implies that there exist two residues with almost the same *?* and $f$ angles. Obviously, the larger the predefined distance e is, the more edge counts will be. How to determine such a predefined distance e is relied on the comparisons among homologous sequences. In other words, it should be pre-computed for each homologous family to find the proper e. Hence, the structurally conserved regions after backbone structure comparison for these two proteins, $p_s$ and $p_t$, can be definitely represented by the graph G = (V, E) where $V = \{\{v^s_i\}, \{v^t_j\}\}$ and $E = \{e_k\}$. The structurally conserved regions between proteins $p_s$ and $p_t$ are demonstrated in Fig. 3 in which two short proteins, $m=11$ and $n=9$, are shown in convenience.

(2) Elimination Policy

From Fig. 3, we can see that several vertices may be incident to a same vertex after step (1). This situation should be ruled out. Here, we use the substitution matrix to compute the mismatch penalty for each possible vertex-to-vertex arrangement. After such a scoring operation, several improper edges will be eliminated. For example, in Fig. 4, both (A) and (B) are the best candidates with highest score 20 if the substitution matrix BLOSUM62 is used. Note that multiple or pairwise sequence alignment also adopts substitution matrix to achieve more reasonable protein residue replacement. But elimination policy here may allow a totally different sequence order as described in section 2.

(3) Adjacency

If there are at least two candidates after step (2), choose those with the highest adjacency density. Such a choice is based on genomics recombination result. In Fig. 4, the structure alignment for (A) will be more possible than (B) since the latter behaves lower adjacency density. However, since the sequence alignment can not be executed in reverse direction, only (B) may be obtained if we perform pairwise sequence alignment between proteins $p_s$ and $p_t$. Hence this algorithm will present more extensive optimal solutions than those by pairwise sequence alignment only.

## 4. Discussion

Here, we only take backbone conformation into account, excluding any side chain. If side chain is also considered, the dimensions for vector $v_i$ will be increased since there are about one to four kinds of dihedral angles ? to be included.

Because the 3D structure has been transformed into the graph G = (V, E), a set of structurally conserved regions should be derived from graph theory. For example, in Fig. 3, the weight for each edge can be measured as the distance $\| v^s_i - v^t_j \|$. Meanwhile, the problem to find the structurally conserved regions is reduced to discover the number of the set of vertices with the minimum distance.

We conclude that sequence alignment is useful in designing evolutionary relationship such as phylogenetic tree. But it may be not sufficient to predict the structurally conserved regions only based on sequence information. On the contrary, in case of more homologous protein structure information obtained in advance, it will get more accuracy to calculate the structurally conserved regions by the comparison of the dihedral angles for the target protein with the well-known tertiary structure database. As a result, if all the structurally conserved regions are orderly collected together like rotamer library, the homologous backbone structure prediction should also be well done.

## 5. Caption

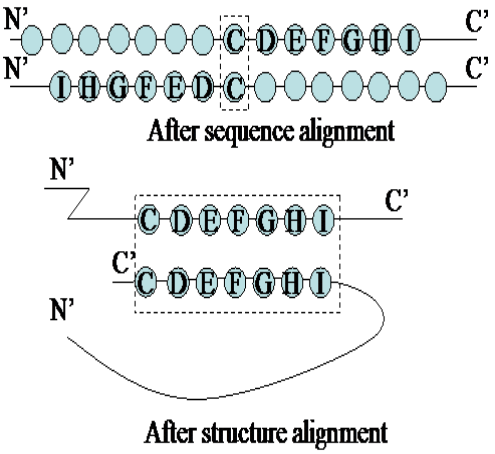


Fig. 1 Residues with reverse sequence order.

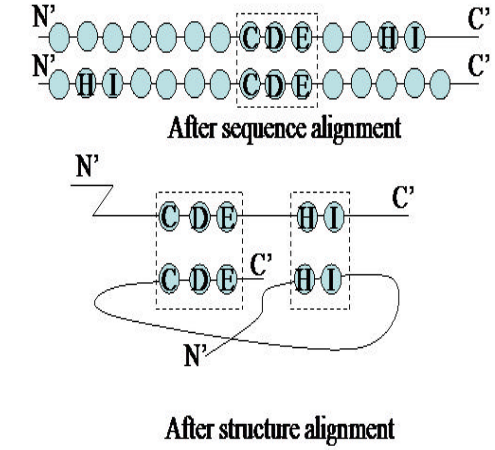


Fig. 2 Residues located distantly with each other.

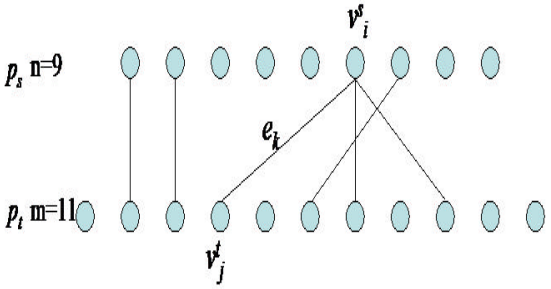


Fig. 3 Demonstration for Relation Graph

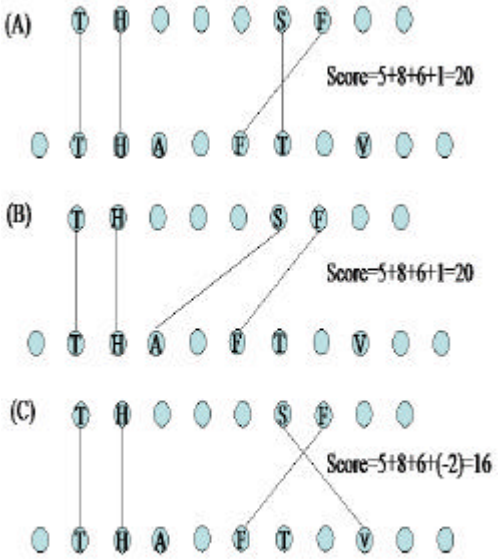


Fig. 4 Demonstration for Elimination Policy

## 6. Reference

[1] Feng, D. F. and Doolittle, R. F. (1987) Progressive sequence alignment as a perquisite to correct phylogenetic trees. *J. Mol. Evol.* 25(4), 351-360.

[2] Taylor, W. R. (1988) A flexible method to align large numbers of biological sequences. *J. Mol. Evol.* 28, 161-169.

[3] Chou, K. C. and Carlacci, L. (1991) Simulated annealing approach to the study of protein structures. *Protein Eng.* 4, 661-667.

[4] Goldberg, D. E. (1989) Genetic algorithms, in *Search, Optimization & Machine Learning*. Addison-Wesley.