Journal of Molecular Structure: THEOCHEM 909 (2009) 102–106

Contents lists available at ScienceDirect

Journal of Molecular Structure: THEOCHEM

journal homepage: www.elsevier.com/locate/theochem

# Use of information discrepancy measure to compare protein secondary structures

Shengli Zhang *, Lianping Yang, Tianming Wang

*School of Mathematical Sciences, Dalian University of Technology, Dalian 116024, PR China*

ARTICLE INFO

*Article history:*
Received 8 May 2009
Received in revised form 26 May 2009
Accepted 30 May 2009
Available online 7 June 2009

*Keywords:*
Protein 3D structures
Transition probability matrix
FDOD function
Similarity tree

ABSTRACT

Protein secondary structure comparison is an important tool to explore and understand the different aspects of protein 3D structures. In this paper, transition probability matrix and structural characteristic vectors of proteins are constructed. Then the FDOD score scheme is developed to measure the similarity and the similarity tree of 20 proteins from four different classes and TOPS strings of the 36 protein chains in the Chew–Kedem dataset are constructed. The result shows that this new approach to measure the similarities between protein secondary structures is simple to implement, computationally efficient and fast.



## 1. Introduction

Understanding protein structure is central to the post-genomic era. A direct and important method to meet this challenge is protein structure. It is well known that protein structure is far better conserved through evolution than protein sequence [1]. That is to say, if similarity between two proteins is detectable at the sequence level, structural similarity can usually be assumed. Moreover, even proteins that have nondetectable sequence similarity may have similar structures, it has been estimated that approximately one-third of all sequences are recognizably related to at least one known protein structure [2–6]. Therefore, structure comparisons are expected to get a more reliable taxonomy,especially for proteins distantly related to each other. Up to now, several methods such as SSAP [7], DALI, CE [8], MAMMOTH [9] and SSM [10] have been developed for this purpose. Structure comparison method will be a useful assessment tool for the protein structure prediction method.

However, the detection of 3D structure similarity presents an enormous computational and theoretical challenge [11,12]. In theory, there is no clear statistical definition of what constitutes an excessive amount of similarity. This is due largely to three circumstances: (1) the range of protein structures appears far more constrained by chemical and physical forces than the range of sequences, (2) there is no definition of an optimal 3D alignment and (3) it is difficult to compare very different protein structures (e.g., all-$\alpha$ versus all-$\beta$). To bypass the difficulty, Przytycka et al. [13] presented a new comparison of protein structure. Instead of utilizing the whole 3D structure, they consider only secondary structures of proteins. In the first step, they reduced a protein sequence to an ordered sequence of its secondary structure elements, i.e., H(Helix), S(Strand) and L(Loop). Then similarity tree of the chosen proteins is got by simply aligning these ordered sequences. It is found that, even at this simple level of reduction, this method can reasonably classify proteins from different SCOP categories. However, their method also suffers from the problems accompanied by sequence alignment, e.g., computational complexity and different sequence lengths. What is more, it is more or less subjective to determine the alignment score matrix, which will seriously affect the final alignment. and Zheng [14] proposed a valid method to compare protein secondary structures based on backbone dihedral angles. Also, some researchers developed graphical techniques to deal with this problem [15–18]. Basically, each molecular sequence or structure is represented as a series of dots in Euclidean space, and some graphical invariants are extracted to characterize the corresponding sequence.

In this paper, a novel approach for protein structure comparison is proposed. The contribution comes from both the representation of protein and score scheme respectively. Firstly, a protein sequence is mapped into a protein secondary structure sequence according to the $\alpha$-Helices, $\beta$-Strands and Coils, then the transition probability matrices and structural characteristic vectors of proteins are constructed. Thirdly, a similarity score called FDOD function (Function of Degree of Disagreement), which is a new measure of information discrepancy, is applied to achieve the comparison of two subsequence distributions of distances. It provides a new measure of protein similarity and has many good mathematical properties. Combining the above phases, a pairwise comparison

* Corresponding author. Tel.: +86 411 8474 9735; fax: +86 411 8470 8354.
*E-mail addresses:* shengli0201@163.com (S. Zhang), wangtm@dlut.edu.cn (T. Wang).

0166-1280/$ - see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.theochem.2009.05.031

algorithm is designed. Finally, the similarity tree of two experiment dataset are constructed to confirm the validity of our method.

## 2. Materials and methods

### 2.1. Construction of transition probability matrices of proteins

Fig. 1 shows the secondary structures of protein, whose PDB code is 1ayd, and it belongs to $\alpha+\beta$ structural classes. In this graph, the secondary structures of a protein are defined by the local back-bone conformation at each position. Secondary structure elements of greatest interest include α-helices(wave)and β-strands(wide arrowhead). They are represented as $H$ and $E$, respectively, in the 1D summary. Remaining positions are represented by $C$ for coil. A secondary structure sequence is a symbolic string composed of three kinds of letters, indicating the helix, strand, and coil, respectively.

A secondary structure sequence [19–21] is a linear sequence defined over state space $S=\{C,H,E\}$, where $H$ represents helix, $E$ represents strand and the rest are represented by $C$(mainly coil and turn). Consider the definition of stochastic process: a stochastic process is a collection $\{X(t)|t\in T\}$ of random variables $X(t)$ defined on the probability space $(\Omega,\Gamma,P)$, where $T$ is called index set, $\Omega$ represents the sample space that is constituted by all the basic events, $\Gamma$ represents the event set that is constituted by all possible events and P represents the probability, which is a function defined over $\Gamma$. Given any $t$, the possible values of $X(t)$ are called the states of the process at $t$. So the secondary structure sequence may be regarded as a realization of a stochastic process. In this stochastic process, the states of the stochastic process are $\{C,H,E\}$ and the index set is a finite ordered sequence of non-negative integer numbers. Then transition probability matrix may be employed to describe a realization of a stochastic process. It records the overall situation that certain state transfers to another state in a realization. The transition probability matrix (TPM) can be defined as follows:

$$\mathrm{TPM}=\begin{pmatrix} P_{HH} & P_{HE} & P_{HC} \\ P_{EH} & P_{EE} & P_{EC} \\ P_{CH} & P_{CE} & P_{CC} \end{pmatrix}$$

They are computed by the following formula:

$$P_{a_ia_j}=\begin{cases} N_{a_ia_j}/\Sigma_{k=1}^{3}N_{a_ia_k} & \Sigma_{k=1}^{3}N_{a_ia_k}\neq 0; \\ 0 & \Sigma_{k=1}^{3}N_{a_ia_k}=0. \end{cases}$$

where $a_i$ represents the $i$th element of state space {H,E,C}; $N_{a_ia_j}$ enumerates the frequency of the incident that letter $a_i$ is followed by letter $a_j$ in a secondary structure sequence.

### 2.2. Construction of structural characteristic vectors of proteins

In order to characterize protein secondary structures numerically, we construct structural characteristic vectors of proteins. The more information the vector extracts, the better the classification result will be. Because both transition probability matrix and the content of elementary structural units are indices from different perspectives, so we integrate them together to define the structural characteristic vector (SCV):

$$\mathrm{SCV}=(P_{HH},P_{HE},P_{HC},P_{EH},P_{EE},P_{EC},P_{CH},P_{CE},P_{CC},m_H,m_E,m_C)$$

where $m_H,m_E$ and $m_C$ represent the content of $H,E$, and $C$, respectively.

The structural characteristic vector characterize secondary structure sequences numerically. They generalize the distribution patterns of elementary structure units in secondary structure sequences and are indices for their corresponding protein structures.

Thus, the structural characteristic vectors can be defined the set of distributions of elements on the function of degree of disagreement (FDOD) in the next section.

### 2.3. FDOD score scheme

Function of Degree of Disagreement (FDOD) is a new measure of information discrepancy [22]. It has been successfully used to measure the discrepancy between DNA sequences and amino acid sequences from different species in the study of phylogeny and prediction of protein structural classes. This measure has a close connection with Shannon entropy, and has many good mathematical characteristics, such as symmetry, boundedness, triangle inequality, and so on. Also this measure is applicable to the multiple sequence comparison [22]. It is a very important property in our study to achieve easily both protein pairwise and multiple structure comparisons.

Given a set of distributions of elements:

$$\begin{aligned} U_1&=(P_{11},P_{21},\ldots,P_{n1}) \\ U_2&=(P_{12},P_{22},\ldots,P_{n2}) \\ &\ldots \\ U_s&=(P_{1s},P_{2s},\ldots,P_{ns}) \end{aligned}$$

where $\sum_{i=1}^{n}P_{ik}=1,k=1,2,\ldots,s$. The FDOD measure is defined as

$$D(U_1,U_2,\ldots,U_s)=\sum_{k=1}^{s}\sum_{i=1}^{n}P_{ik}\cdot\log P_{ik}/\left(\sum_{k=1}^{s}P_{ik}/s\right)$$

where $0\cdot\log 0=0$ and $0\cdot\log 0/0=0$ are defined. $D(U_1,U_2,\ldots,U_s)$ denotes a measure of discrepancy among distributions.

For any two distributions $U_1=(P_{11},P_{21},\ldots,P_{n1})$ and $U_2=(P_{12},P_{22},\ldots,P_{n2})\in\Gamma_n$, the FDOD measure reduces into:

$$D(U_1,U_2)=\sum_{i=1}^{n}P_{i1}\cdot\log(2P_{i1}/(P_{i1}+P_{i2}))+\sum_{i=1}^{n}P_{i2}\cdot\log(2P_{i2}/(P_{i1}+P_{i2}))$$

It has been proved that the FDOD measure can be a distance measure [23]. So we substitute $U$ with $SCV$, then pairwise distance matrix is got on the basis of a measure of information discrepancy.

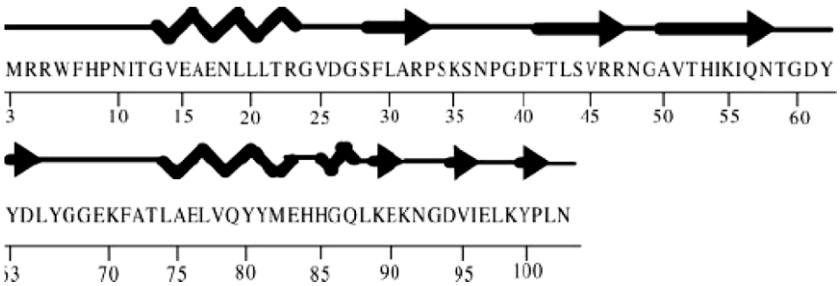


**Fig. 1.** The secondary structure of the protein 1ayd.