Available online at www.sciencedirect.com
ScienceDirect
Journal of Theoretical Biology 250 (2008) 382–387
Journal of Theoretical Biology
www.elsevier.com/locate/yjtbi

Comparison of protein secondary structures based on backbone dihedral angles

Jun Wang[a,b], Xiaoqi Zheng[a,*]

[a]Department of Applied Mathematics, Dalian University of Technology, Dalian 116024, China
[b]College of Advanced Science and Technology, Dalian University of Technology, Dalian 116024, China

Received 31 July 2007; received in revised form 12 October 2007; accepted 12 October 2007
Available online 17 October 2007

Abstract

In this article, we propose a relatively similar measure to compare protein secondary structures. We first transform a protein secondary structure into a special sequence representation (angle sequence) based on a partition of the backbone ϕ,ψ-space. Then, pairwise sequence distance is evaluated on the basis of a symbolic sequence complexity. To illustrate our approach, we construct the similarity tree of 24 proteins from PDB.

Keywords: Similarity tree; Protein taxonomy; Sequence complexity

1. Introduction

The comparison of protein structures has been an extremely important problem in structural and evolutionary biology ever since the first few protein structures became available. It is known that protein structure is far better conserved through evolution than protein sequence (Chothia and Lesk, 1986). 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 (Fischer and Eisenberg, 1997; Huynen et al., 1998; Jones, 1999; Rychlewski et al., 1998). Therefore, structure comparisons are expected to get a more reliable taxonomy, especially for proteins distantly related to each other. Till now, several methods such as SSAP (Taylor and Orengo, 1989), DALI, CE (Shindyalov and Bourne, 1998), MAMMOTH (Ortiz et al., 2002) and SSM (Krissinel and Henrick, 2004) have been developed for this purpose.

However, the detection of 3D structure similarity presents an enormous computational and theoretical challenge (Gibrat et al., 1996). Actually, by whatever metric is chosen, the similarity of 3D structure means the similarity of the relative spatial orientation of many points drawn from each structure, for example, the coordinates of C_α atoms in the polypeptide backbone. But there are a very large number of ways in which one could match backbone atoms from any two proteins, and brute force computation is totally infeasible with today's computers. 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 specify a "random" protein structure, and it is difficult to compare very different protein structures (e.g., all-α versus all-β).

To bypass the difficulty, Przytycka et al. (1999) presented a new comparison of protein structure. Instead of utilizing the whole 3D structure, they consider only secondary structures of proteins (backbone dihedral angles, explicitly). In the first step, they reduced a protein sequence

*Corresponding author. Tel.: +86 411 84749342; fax: +86 411 84706100.
E-mail address: xiaoqizheng@yahoo.com.cn (X. Zheng).

0022-5193/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jtbi.2007.10.013

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. Alternatively, some researchers developed graphical techniques to deal with this problem (Liao et al., 2006a, b; Randić et al., 2005, 2006). Basically, each molecular sequence/structure is represented as a series of dots in Euclidean space, and some graphical invariants are extracted to characterize the corresponding sequence.

It is worth mentioning whether or not protein secondary structure determines the 3D fold is still controversial to modern biologists. Both experiment (Minor and Kim, 1996) and theory (Yee et al., 1994) suggest that tertiary structure may precede secondary structure during the folding of a protein. But experiments of Przytycka et al. prove that secondary structure alone may be sufficient to recognize the tertiary fold.

Motivated by Przytycka et al., in this article we present a new method to evaluate the distance between two protein structures. Based on a partition of the backbone ϕ, ψ-space, protein secondary structures are approximated as symbol sequences with nine letters. Then pairwise distance matrix is got on the basis of a symbolic sequence complexity (Lempel–Ziv (LZ) complexity). Finally, a taxonomy of 24 proteins is constructed to confirm the validity of our method.

2. Materials and methods

2.1. Materials

A representative set of proteins is selected from the Protein Data Bank (PDB) in Table 1. In order to highlight our structure-based method, any pair of these proteins is guaranteed to have an alignment sequence identity less than 30%. In this case, however, weak sequence homology can be detected, and sequence-based methods can hardly get reliable results.

2.2. Angle sequence for protein secondary structure

Dihedral angles in proteins are of significant importance because they fully determine the backbone configurations of a protein, so in most cases, they play a key role in defining or ''tightening'' the protein secondary structure. In native folded protein, there exists a high preference for ϕ and ψ in three major types of secondary structure (α-helix, β-structure and β-turn) (Fig. 1(a)). This preference for backbone dihedral angles arises from a combination of steric effects, both within individual amino acid residues and between side-chains of different residues, and of secondary structure interactions such as hydrogen bond formation (Morris et al., 1992). Consequently, the backbone dihedral angle is a crucial measure in attempting to

Table 1
Proteins used in this article

No.	PDB ID	Taxonomy	Family	Class	Length (nt)
1	1eca	*Chironomus thummi thummi*	Heme bingding	α protein	136
2	1mbd	*Physeter catodon*	Heme bingding	α protein	153
3	1hbg	*Glycera dibranchiata*	Heme bingding	α protein	147
4	1mba	*Aplysia limacina*	Heme bingding	α protein	146
5	3c2c	*Rhodospirillum rubrum*	Monodomain cytochrome c	α protein	112
6	351c	*Pseudomonas aeruginosa*	Monodomain cytochrome c	α protein	82
7	1fha	*Homo sapiens*	Ferritin	α protein	170
8	1rci	*Rana catesbeiana*	Ferritin	α protein	171
9	1cd8	*Homo sapiens*	V-set domain, immunoglobulin	β protein	114
10	2rhe	*Homo sapiens*	V-set domain, immunoglobulin	β protein	114
11	1cdb	*Homo sapiens*	V-set domain, immunoglobulin	β protein	105
12	1cdc	*Rattus norvegicus*	V-set domain, immunoglobulin	β protein	193
13	1plc	*Populus nigra*	Plastocyanin/azurin	β protein	99
14	1aaj	*Paracoccus denitrificans*	Plastocyanin/azurin	β protein	105
15	2bat	*Unidentified influenza virus*	Sialidase	β protein	388
16	2sim	*Salmonella typhimurium*	Sialidase	β protein	381
17	5p21	*Homo sapiens*	G protein	α/β protein	166
18	1etu	*Escherichia coli B*	G protein	α/β protein	178
19	3dfr	*Lactobacillus casei*	Dihydrofolate reductase	α/β protein	162
20	8dfr	*Gallus gallus*	Dihydrofolate reductase	α/β protein	186
21	1onc	*Rana pipiens*	Ribonuclease A	$\alpha+\beta$ protein	104
22	7rsa	*Bos taurus*	Ribonuclease A	$\alpha+\beta$ protein	124
23	3il8	*Homo sapiens*	Interleukin 8-like chemokine	$\alpha+\beta$ protein	68
24	1ctf	*Escherichia coli*	Ribosomal protein L7/12	$\alpha+\beta$ protein	68