doi:10.1016/j.jmb.2008.12.044 J. Mol. Biol. (2009) 387, 431–450

JMB

Available online at www.sciencedirect.com

ScienceDirect

ELSEVIER

Structural Alphabets for Protein Structure Classification: A Comparison Study

Quan Le[1], Gianluca Pollastri[1] and Patrice Koehl[2]*

[1]*Complex and Adaptive Systems Laboratory, School of Computer Science and Informatics, University College Dublin, Dublin, Ireland*

[2]*Department of Computer Science and Genome Center, University of California, Davis, Davis, CA 95616, USA*

Received 21 July 2008; received in revised form 16 December 2008; accepted 17 December 2008
Available online
25 December 2008

Finding structural similarities between proteins often helps reveal shared functionality, which otherwise might not be detected by native sequence information alone. Such similarity is usually detected and quantified by protein structure alignment. Determining the optimal alignment between two protein structures, however, remains a hard problem. An alternative approach is to approximate each three-dimensional protein structure using a sequence of motifs derived from a structural alphabet. Using this approach, structure comparison is performed by comparing the corresponding motif sequences or structural sequences. In this article, we measure the performance of such alphabets in the context of the protein structure classification problem. We consider both local and global structural sequences. Each letter of a local structural sequence corresponds to the best matching fragment to the corresponding local segment of the protein structure. The global structural sequence is designed to generate the best possible complete chain that matches the full protein structure. We use an alphabet of 20 letters, corresponding to a library of 20 motifs or protein fragments having four residues. We show that the global structural sequences approximate well the native structures of proteins, with an average coordinate root mean square of 0.69 Å over 2225 test proteins. The approximation is best for all α-proteins, while relatively poorer for all β-proteins. We then test the performance of four different sequence representations of proteins (their native sequence, the sequence of their secondary-structure elements, and the local and global structural sequences based on our fragment library) with different classifiers in their ability to classify proteins that belong to five distinct folds of CATH. Without surprise, the primary sequence alone performs poorly as a structure classifier. We show that addition of either secondary-structure information or local information from the structural sequence considerably improves the classification accuracy. The two fragment-based sequences perform better than the secondary-structure sequence but not well enough at this stage to be a viable alternative to more computationally intensive methods based on protein structure alignment.

© 2009 Elsevier Ltd. All rights reserved.

Keywords: protein structure; structural alphabet; structure classification; protein sequence comparison; sequence feature space

Edited by M. Levitt

Introduction

There is a clear understanding in biology that all cellular functions are deeply connected to the shape of their molecular actors. This is especially true for proteins, whose functions are directly related to their three-dimensional (3D) structures.[1–4] In the hope of deciphering the rules that define the relationships between structure and functions, large-scale experimental projects are performed to provide maps of the genetic information of different

*Corresponding author. E-mail addresseses: quan.le@ucd.ie; gianluca.pollastri@ucd.ie; koehl@cs.ucdavis.edu.

Abbreviations used: 3D, three-dimensional; NS, native sequence; SSES, secondary-structure element sequence; LFS, local fragment sequence; GFS, global fragment sequence; SVM, support vector machine; HMM, hidden Markov model; cRMS, coordinate root mean square; MDS, multidimensional scaling; AIS, average intercluster separation; ROC, receiver operating characteristic; TP, true positive; FP, false positive; CPU, central processing unit.

organisms, including the human genome[5,6] (mostly in the form of genetic sequences of proteins), to derive as much structural information as possible on the products of these genes, and to relate these structures to the function of the corresponding proteins. These are the different "-omics" projects, genomics, structural genomics,[7] and functional genomics,[8] to name a few. While these studies are generating a wealth of information, stored into databases, the key to their success lies in our ability to organize and analyze this information, that is, in our ability to classify proteins, based on their sequences, structures, and/or functions, and to connect these classifications (for reviews, see Refs. [9,10]). In this article, we focus on the effort of organizing protein structures.

It is currently easier to detect that two proteins share similar functions based on their structures rather than on their sequences. This was observed as early as in 1960, when Perutz *et al.* showed that myoglobin and hemoglobin, the first two protein structures to be solved at atomic resolution using X-ray crystallography, have similar structures even though their sequences differ.[11] These two proteins are functionally similar, as they are involved in the storage and the transport of oxygen, respectively. Since then, there has been a continued interest in finding structural similarities between proteins, in the hope of revealing shared functionality that could not be detected by sequence information alone. The result of this interest is the development of systems for classification of protein structures that identify and group proteins sharing the same structure so as to reveal evolutionary relationships. Currently, there are three main protein structure classification schemes: SCOP,[12] CATH,[13] and DALI.[14]

Central to any classification scheme is the concept of similarity and its quantification. A measure of similarity is required to generate the initial classification of the data, as well as to identify the class to which any new data would belong. Defining a similarity measure for protein structures is a difficult problem, leading to discrepancies between the different classification schemes. Protein structure similarity is most often detected and quantified by a protein structure alignment program. Although an approximate optimal solution of the structural alignment problem exists,[15] it is computationally too expensive to be of practical interest. All methods available to date are heuristic and, consequently, at best, suboptimal. Many evaluations of structural alignment methods are available.[16–19] These studies usually conclude that an optimal solution to this problem that is fast and reliable and therefore appropriate for classification still needs to be defined. As a consequence, there is a significant interest in developing alternative approaches to protein structure alignment for measuring similarities (for a recent review, see Ref. [10]).

Finding the (sub) optimal alignment between two protein structures is a hard problem as the rotation and translation of one of the two structures with respect to the other must be found in addition to the alignment itself. By converse, finding the optimal alignment between two protein sequences is a much easier problem, as it can always be solved using dynamic programming, so long as a satisfactory substitution model is available. If it was possible to translate faithfully a protein structure into a string of letters, protein structure comparison would therefore become much easier. This idea of representing structures as a string of letters is in fact grounded in the observation that recurrent, regular structural motifs exist at all levels of organization of protein structures. This was first observed by Corey and Pauling[20,21] and later refined into the concept of protein secondary structures. Although the latter can be predicted with high accuracy (>80%), the description of a protein in terms of its secondary structures is not sufficient to capture accurately its 3D geometry. To overcome this limitation, several studies have focused on defining libraries of fragment representatives from which complete protein structures can be modeled with adequate accuracy.[22–29] In these approaches, protein structures are represented as a series of overlapping fragments, each labeled with a symbol, defining a structural alphabet for proteins. The fragments are chosen such as to provide either the best local fits to segments of the protein structure or the best global fit to the entire protein structure, resulting in two types of structural sequences, namely, local or global.[27] Current applications of these structural alphabets include protein structure modeling and in particular decoy generation,[30] local structure prediction,[24,31–34] the reconstruction of a full-atom representation of the protein from the knowledge of the positions of its C^{α} only,[35,36] identification of structural motifs,[37,38] analysis of protein–protein interactions,[39] and protein structure comparison and protein structure database search.[29,40–46] We are interested in an extension of the latter, that is, to the application of structural alphabets to the problem of protein structure classification.

In this article, we focus on the information content of sequences of proteins, in the context of structure classification (fold classification). More specifically, we compare the performance of five different classifiers, each tested with four different sequence representation of proteins: the native (amino acid) sequences (NSs), the secondary-structure element sequence (SSES), and two fragment-based structural sequences, namely, a local fragment sequence (LFS) and a global fragment sequence (GFS) derived from a library of 20 fragments having four residues. We show that LFS, GFS, and SSES always outperform the NSs and that GFS and LFS perform statistically significantly better than the SSES when adopted in combination with kernel-based, support vector machine (SVM)-based, and hidden Markov model (HMM)-based classifiers.

Results

With the number of known protein structures in the Protein Data Bank[47] growing exponentially, the need for reliable, automatic structure comparison

Download English Version:

https://daneshyari.com/en/article/2186604

Download Persian Version:

https://daneshyari.com/article/2186604

Daneshyari.com