Computers in Biology and Medicine 59 (2015) 64–72

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

Reducing dimensionality in remote homology detection using predicted contact maps

Oscar Bedoya*, Irene Tischer

School of Computer Science and Systems Engineering, Universidad del Valle, Cali, Colombia

ARTICLE INFO

Article history:
Received 22 October 2014
Accepted 22 January 2015

Keywords:
Classification
Physicochemical properties
Remote homology detection
SCOP family
3D structure models

ABSTRACT

In this paper, a new method for remote protein homology detection is presented. Most discriminative methods concatenate the values extracted from physicochemical properties to build a model that separates homolog and non-homolog examples. Each discriminative method uses a specific strategy to represent the information extracted from the protein sequence and a different number of indices. After the vector representation is achieved, support vector machines (SVM) are usually used. Most classification techniques are not suitable in remote homology detection because they do not address high dimensional datasets. In this paper, we propose a method that reduces the high dimensionality of the vector representation using models that are defined at the 3D level. Next, the models are mapped from the protein primary sequence. The new method, called remote-C3D, is presented and tested on the SCOP 1.53 and SCOP 1.55 datasets. The remote-C3D method achieves a higher accuracy than the composition-based methods and a comparable performance with profile-based methods.

1. Introduction

Remote homology detection is a fundamental problem in bioinformatics. A remote homolog of a protein *P* is another protein *Q* that does not exhibit high sequence similarity but is functionally and structurally related. This type of research is considered a fundamental step in biomedical applications such as drug discovery [1,2], where there is a need to identify proteins that share common functions given a specific protein sequence. Although the formal definition of remote homology detection refers to protein sequences with less than 25% sequence identity, called the "twilight zone" in Homaeian et al. [3] and Huang and Bystroff [4], the problem can also be defined using the SCOP hierarchy. Considering the four levels of the SCOP hierarchy (i.e., class, fold, superfamily and family), the remote homologs of a protein *P* in family F are proteins in the same superfamily of P that do not belong to F.

Several methods have been proposed to determine remote homology [2,5–18]. Recent strategies include discriminative methods, which build a model considering both positive (remote) and negative (non-remote) examples. They rely on a combination of features that can discriminate protein families. SVM I-sites [5] is a discriminative method. In SVM I-sites, every subfragment of an unknown target sequence is submitted to the log-odd matrix representing each I-site. The I-sites library is a set of short sequence patterns (profiles) that correlate strongly with protein three-dimensional structure elements. The similarity scores of different clusters of I-sites are mapped to a range of comparable values by using a confidence curve. A confidence curve maps similarity scores to the probability of the correct local structure based on a jack-knife test.

Remote homology can be detected by sequence composition-based methods, which are based on using subsequences, motifs or word similarity from protein sequences to extract features that help in discriminating protein families. There are sequence composition-based methods that incorporate physicochemical properties of amino acids [7–10]. SVM-RQA [7] is a discriminative method that uses a recurrent quantification analysis to measure the similarity between two proteins. The recurrent quantification analysis is a method that allows detecting recurrence patterns along the protein sequence. A protein is represented by 10 values extracted from an embedding matrix that is part of a recurrence quantification analysis. An embedding matrix can be calculated for each physicochemical property. SVM-RQA uses 480 indices or physicochemical properties, obtaining a total of $10 \times 480 = 4800$ values in its vector representation. Another discriminative method is SVM-PCD [8], which proposes to represent a protein by 18 values extracted from the distribution of physicochemical properties scores. Each distribution is calculated from the 4-mers of the primary sequence when a single physicochemical property is considered; a 4-mer is the average of the physicochemical values in a 4-size window. Webb-Robertson et al. [8] propose SVM-PCD (61), SVM-PCD (181), and SVM-PCD (531), which uses 61, 181, and

* Corresponding author. Tel.: +57 2 3212100x2781; fax: +57 2 3212100.
E-mail address: oscar.bedoya@correounivalle.edu.co (O. Bedoya).

http://dx.doi.org/10.1016/j.compbiomed.2015.01.020
0010-4825/

531 physicochemical properties, respectively. Thus, a total of $61 \times 18 = 1098$, $181 \times 18 = 3258$, and $531 \times 18 = 9558$ values are used in the vector representation. Identifying remote homologs is performed by using the support vector machines (SVM) technique because it is not affected by the high dimensionality of the vector representation of proteins (usually 4200 values or more for each protein). Building an SVM depends on the number of samples in the training dataset and not on its dimensionality [19]. Most of the classification techniques (i.e., Bayes classifiers, neural networks, decision trees) are not suitable when they are used on high dimensional data. For instance, calculating the probabilities when a Bayes classifier is being built from high dimensional data makes that classification technique unfeasible [20]. Using other classification techniques rather than support vector machines could be achieved by reducing the dimensionality of the protein representation in the remote homology detection problem.

Remote homology can also be detected by using generative methods, which are based on building a statistical model for each family or superfamily. Hidden Markov Models (HMM) and profiles of protein families are frequently used [14,15]. Secondary structure elements have also been considered in remote homology detection [17,18]. Kumar and Cowen [17] focus on detecting remote homology for β-structural motifs by training a HMM model with sequences based on pairwise dependencies of β-sheet hydrogen bonding. Profile-based methods are another kind of strategies that allow detecting remote homology. Profile-based methods [21–23] use an alignment of each protein against a non-redundant database to obtain a profile. The evolutionary information extracted from profiles allows detecting remote homologs in a more accurate way than composition-based methods.

In this paper, we propose a new method that reduces the high dimensionality of the vector representation in remote homology detection by using models that are defined at the 3D level and thus are highly structurally and functionally related. Then, the 3D models are mapped from the protein primary sequence. We propose to address the problem of remote homology detection by correlating 3D structure models and primary sequence. The new method, called remote homology detection by the correlation of 3D models (remote-C3D), is presented and tested on two different datasets, SCOP 1.53 and SCOP 1.55.

In the following section every step in the remote-C3D method is explained in detail. In Section 3, the results are given for both the SCOP 1.53 benchmark and the SCOP 1.55 dataset. We focus the evaluation of the remote-C3D method on discovering the effect of reducing the dimensionality of the vector representation in remote homology detection and also on obtaining the accuracy of a remote homology detection method that is based on 3D structural models. Finally, the conclusions of the research are presented in Section 4.

2. Methods

In this section, we explain in detail every step in the remote-C3D method. The remote-C3D method includes obtaining the 3D models, predicting a contact map from the primary sequence, calculating the count vector, and building a classifier for each SCOP family. The general overview of the remote-C3D method is shown in Fig. 1.

2.1. Obtaining the 3D models

The first step in the remote-C3D method is obtaining 3D models from the contact map representation of protein in a given dataset. A distance matrix of a protein is a square matrix containing the Euclidean distances between all pairs of Cα atoms in the protein. A contact map is achieved by discretizing the distance matrix. Different thresholds can be used to determine when two residues are in contact. In this research, we assume that two residues are in contact when the Euclidean distance between the corresponding Cα atoms is less than or equal to 8.0 Angstroms. A cutoff distance of 8.0 Angstroms has been considered a standard threshold for most of the current contact maps prediction programs [24–26].

Although contact maps of any two proteins are different, there are specific areas in the contact maps that can be recognized as patterns even in different domains. In fact, Choi et al. [27] found that a set of 100 representative 10×10 submatrices extracted from the distance matrices are able to discriminate domains in the protein structure comparison problem, which is related to classifying a protein into the correct SCOP classification at the class, fold, superfamily, and family levels. Choi et al. [27] established that the 100 representative submatrices or models reflecting local structural features, and combinations of these models can be used to reconstruct the original distance matrices. In addition, Choi et al. [27] demonstrated that even though there are millions of different submatrices in all proteins, most of the submatrices are common, and thus, a finite number of models can be sufficient to represent observed interactions in the distance matrix.

If the 3D information is available then solving the remote homology detection problem is very easy; otherwise it becomes a complicated task. Therefore, the remote-C3D method predicts contact maps from the primary sequence alone. Even though

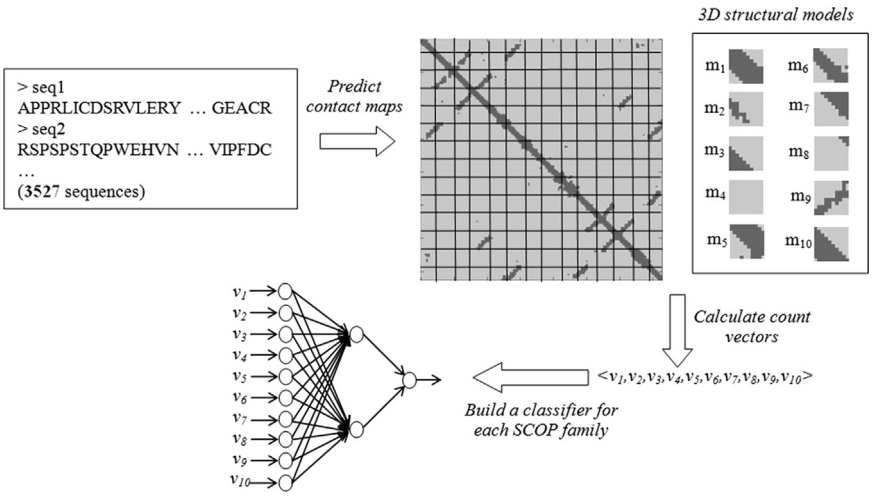

Fig. 1. General overview of remote-C3D.