Available online at www.sciencedirect.com
SCIENCE DIRECT
Computational Biology and Chemistry
Computational Biology and Chemistry 30 (2006) 268–279
www.elsevier.com/locate/compbiolchem

Identifying the interacting positions of a protein using Boolean learning and support vector machines

Anshul Dubey [a], Matthew J. Realff [a,*], Jay H. Lee [a], Andreas S. Bommarius [a,b,c]

[a] School of Chemical and Biomolecular Engineering, 311 Ferst Drive, Atlanta, GA 30332, United States
[b] School of Chemistry and Biochemistry, Georgia Institute of Technology, United States
[c] Parker H. Petit Institute of Bioengineering and Bioscience, Georgia Institute of Technology, United States

Received 16 August 2005; received in revised form 5 April 2006; accepted 6 April 2006

Abstract

It is known that in the three-dimensional structure of a protein, certain amino acids can interact with each other in order to provide structural integrity or aid in its catalytic function. If these positions are mutated the loss of this interaction usually leads to a non-functional protein. Directed evolution experiments, which probe the sequence space of a protein through mutations in search for an improved variant, frequently result in such inactive sequences. In this work, we address the use of machine learning algorithms, Boolean learning and support vector machines (SVMs), to find such pairs of amino acid positions. The recombination method of imparting mutations was simulated to create in silico sequences that were used as training data for the algorithms. The two algorithms were combined together to develop an approach that weighs the structural risk as well as the empirical risk to solve the problem. This strategy was adapted to a multi-round framework of experiments where the data generated in the present round is used to design experiments for the next round to improve the generated library, as well as the estimation of the interacting positions. It is observed that this strategy can greatly improve the number of functional variants that are generated as well as the average number of mutations that can be made in the library.

Keywords: Interacting position; OCAT; Directed evolution; Support vector machines; Boolean function; Recombinations

1. Introduction

The mapping of a protein's sequence to its function is one of the most challenging problems of protein engineering. This mapping is provided physically by the 3D structure into which the amino acids fold and dynamics of conformational changes that occur with target molecules. There have been recent advances through rational design of sequences for which the 3D structure and protein function to structure map is well known (Kuhlman et al., 2003). However, for many classes of proteins the 3D structure, the structure to function map, and conformational dynamics are unknown. Furthermore, even armed with this analytical knowledge, the protein engineer still has a formidable synthetic problem of finding sequences that perform well. This synthetic task is of considerable complexity due to the size of the sequence space that must be searched. In the absence of this knowledge, the current paradigm is to search randomly through sequence space to find improvements, and then to perform successive rounds of hill-climbing using these improvements.

Directed evolution (DE) is a process in which mutations are made, usually at random, in an existing protein sequence in search for a desired property (Chen and Arnold, 1993; Stemmer, 1994). Even after considerable advances in generation and screening of these variants of a protein (Lin and Cornish, 2002; Daugherty et al., 2000; Neylon, 2004; Petrounia and Arnold, 2000), only a small fraction of all the possible sequences (20^L distinct sequences for a protein of length L) can be generated. Moreover, due to the complex nature of the sequence to function map, a majority of the mutations lead to inactive or unfolded proteins. It is known, however, that only a small fraction of the amino acids present in a protein contribute significantly to the protein's properties (Huang et al., 1996). These circumstances provide the motivation to elucidate the importance of each position as well as the amino acid in that position, so that the

* Corresponding author. Tel.: +1 7032927081.
E-mail address: mrealff@chbe.gatech.edu, matthew.realff@chbe.gatech.edu (M.J. Realff).

1476-9271/$ – see front matter © 2006 Published by Elsevier Ltd.
doi:10.1016/j.compbiolchem.2006.04.001

experiments can be conducted more effectively to increase the chances of success.

It has been shown that certain amino acid residues or positions in a protein sequence contribute significantly to the function or its structural stability compared to others (Chen and Arnold, 1993; Huang et al., 1996; Dubey et al., 2005). Apart from single amino acid residues, there also exist pairs of residues that interact with each other in a protein's three-dimensional structure (Meyer et al., 2003; Saraf et al., 2004). Such residues can be identified by applying the concept of feature selection in machine learning to the data that are generated during DE experiments. To create active variants of a protein it is desirable that such residues and residue pairs be preserved. This will increase the probability of finding a variant with an improved function since a larger fraction of the total number of sequences generated will be functional. Also, since a larger number of crossovers can be made without deactivating the protein, more diversity can be achieved.

Dubey et al. (2005) proposed a support vector machines (SVMs) based algorithm to identify the individual positions of a protein, which when mutated can lead to loss of its function. The sequences of active and inactive variants generated during DE, using mutagenesis or recombination, were used as training data. However, the case of interactions between residues was not addressed. Saraf and Maranas (2003) used the structure of a protein to find the pairs that can have a possible interaction. Two parents in a recombination experiment having the same structure (this was assumed even if the structure for one of the parents was not available), and residue positions that are close together in both parents but have different residues, were identified. This approach has its limitations because of its dependence on the availability of three-dimensional structure and the fact that the interactions cannot be verified without doing additional experiments.

Meyer et al. (2003) conducted a similar study based on the structure of the protein. They suggested a SCHEMA guided approach that yields blocks of protein structure, which when swapped between parents result in the minimum amount of disruptions between possibly interacting positions. Experimental verification was provided by shuffling TEM-1 and PSE-4 β-lactamase at 13 selected sites. Based on sequencing a few chimeras, it was shown that, in general, the ones retaining most of the parental activity had lower disruptions.

Support vector machines (SVMs) (Vapnik, 1995; Christianini et al., 2000) are a recently developed learning algorithm that is based on the statistical learning theory (SLT). Data are mapped from the input space to the higher dimensional feature space, using a kernel, so that a linear function can be fitted. The linear function is obtained by minimizing the upper bound on the generalization error. The number of parameters involved in this function is equal to the size of the data set used for training, which avoids over-fitting. SVMs have been widely applied in the field of pattern recognition, for example, face recognition (Deniz et al., 2003; Sanchez and David, 2003). They have also been used for the secondary structure prediction of proteins (Ward et al., 2003; Kim and Park, 2003).

Boolean functions (Schneeweiss, 1989) are a standard representational tool for computer and engineering applications. Given a set of examples in Boolean form (0's and 1's), which are classified as either positive or negative, a Boolean learning algorithm can establish a set of rules or Boolean expressions, which classify all the examples correctly. The primary algorithm used to establish such a Boolean function from the data is called OCAT (One Clause At a Time), which is either based on the branch-and-bound algorithm (Triantaphyllou, 1994) or certain heuristics (Deshpande and Triantaphyllou, 1998). The problem of designing examples that can improve the estimation of the Boolean function, which is known as the guided learning approach has also been investigated (Triantaphyllou and Soyster, 1996; Sanchez et al., 2002).

In the next two sections the basics of SVM and Boolean function learning will be reviewed, and the opportunity for their combination identified, which will be presented in Section 4. In Section 5, the problem of identifying the interacting positions in a protein will be posed as a classification problem to be solved by machine learning.

2. Support vector machines (SVMs)

SVM learning is a highly effective non-linear classification and regression algorithm that are based on statistical learning theory (Vapnik, 1995). Instead of using a non-linear function for classification in the input space, the data are transformed into a feature space where the data are linearly separable (Fig. 1). A linear classification algorithm is used to find an optimum hyperplane that separates the data in that space.

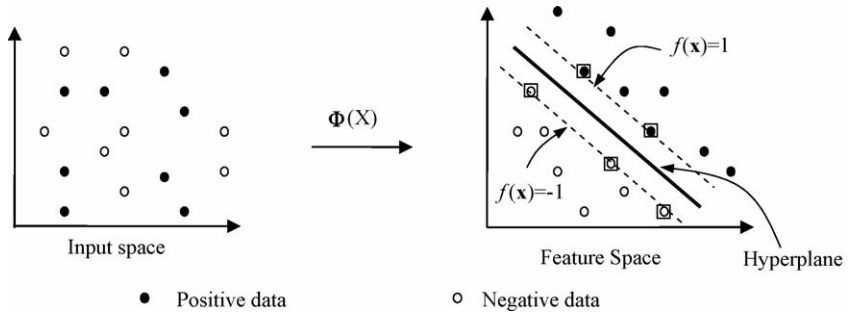

Fig. 1. An illustration of the transformation to the feature space. (●) Positive data; (○) negative data.