Substitution Matrix Optimisation for Peptide Classification

David C. Trudgian and Zheng Rong Yang

School of Engineering, Computer Science and Mathematics,
University of Exeter, Exeter, EX4 4QF, UK
{d.c.trudgian,z.r.yang}@ex.ac.uk

Abstract. The Bio-basis Function Neural Network (BBFNN) is a novel neural architecture for peptide classification that makes use of amino acid mutation matrices and a similarity function to model protein peptide data without encoding. This study presents an Evolutionary Bio-basis network (EBBN), an extension to the BBFNN that uses a self adapting Evolution Strategy to optimise a problem specific substitution matrix for much improved model performance. The EBBN is assessed against BBFNN and multi layer perceptron (MLP) models using three datasets covering cleavage sites, epitope sites, and glycoprotein linkage sites. The method exhibits statistically significant improvements in performance for two of these sets.

1 Introduction

1.1 Background

The study of interactions between proteins within the cell is a major topic of systems biology. Such interactions commonly depend on the successful recognition of suitable functional sites which support them. This recognition process uses the information contained in the 3D conformations and primary structures of the proteins involved. There are many kinds of functional sites which have been examined, including those for protease cleavage, glycosylation, phosphorylation, acetylation and enzymatic catalysis.

Work on the identification of functional sites typically uses sets of fixed length short peptides (short amino acid chains from a protein). These peptides are classified as either functional or non-functional by the trained system. For cases such as protease cleavage sites, the actual functional site will be between two of the residues. In post translational modifications, such as phosphorylation, the functional site is a single residue within the peptide.

A large number of techniques have been applied to functional site recognition. These can be broadly grouped into frequency based, rule based, statistical modelling, and machine learning techniques. Frequency based techniques were the first to be applied using computers and, along with rule based approaches, are simple to interpret. Statistical techniques such as hidden Markov Models (HMMs) and machine learning approaches such as neural networks can attain higher prediction accuracy in many cases.

E. Marchiori, J.H. Moore, and J.C. Rajapakse (Eds.): EvoBIO 2007, LNCS 4447, pp. 291–300, 2007.

1.2 Amino Acid Substitution Matrices

The Bio-basis function neural network (BBFNN), defined in [1], makes use of an amino acid substitution matrix to calculate similarity values between input and Bio-basis peptides. These matrices, developed by Dayhoff and others [2], estimate the probability of mutations between amino acids for a given evolutionary distance. For a set of protein sequences, odds values are calculated for the observed probability of mutation from one amino acid into the other, divided by expected mutation rate (the product of the frequencies of these amino acids i and j). These values are presented in logarithmic form such that the calculations for sequence similarity are additive rather than multiplicative:

$$M_{i,j} = \log_2 \left(\frac{q_{i,j}}{p_i p_j} \right)$$

The selection of a substitution matrix for use with a BBFNN is a difficult decision which can influence performance since it alters the transformation from input into feature space. Currently a trial and error approach is adopted in which performance statistics using a variety of standard matrices are compared, the best then being chosen. This is a time consuming process; the current implementation of the BBFNN includes 15 matrices, excluding variation of evolutionary distances.

The PAM, BLOSUM[3], GONNET [4] and other substitution matrices are constructed using all available sequence data, giving rise to probabilities of mutation averaged across all species, protein functions and both functional and non-functional regions. Whilst this generality is a desirable trait for common sequence searching and alignment tasks it is questionable whether it is appropriate for discriminatory use in a BBFNN applied to a specific classification task. Amino acid frequencies and mutations can vary greatly from the average for all sequences. It is reasonable therefore to expect that problem specific matrices could improve BBFNN classification results on the specific datasets by representing the 'closeness' of amino acids in relation to maintaining a required protein primary sequence structure, e.g. for enzyme catalytic activity, or 3-D formation e.g. for protein disorder prediction.

Generating problem specific matrices is certainly possible by limiting the data from which a matrix is produced to the type of peptide we are trying to classify, or transforming the target frequencies of a general matrix [5]. It is reasonable to expect that such a matrix would result in higher Bio-basis scores for functional sites as it would represent the conserved mutations of these sites. PHAT, a transmembrane-specific matrix, was shown to outperform the general matrices in transmembrane specific searches [6].

An alternative to working forward from sequence data to create a problem specific matrix is to start with a random or standard matrix and work backwards from classifier results to optimise the matrix. Research is ongoing into two such methods of optimisation. The first, which will be discussed in this paper, uses an evolutionary computation approach.