Simplifying Amino Acid Alphabets Using a Genetic Algorithm and Sequence Alignment*

Jacek Lenckowski and Krzysztof Walczak

Institute of Computer Science, Warsaw University of Technology
ul. Nowowiejska 15/19, 00-665 Warszawa, Poland

Abstract. In some areas of bioinformatics (like protein folding or sequence alignment) the full alphabet of amino acid symbols is not necessary. Often, better results are received with simplified alphabets. In general, simplified alphabets are as universal as possible. In this paper we show that this concept may not be optimal. We present a genetic algorithm for alphabet simplifying and we use it in a method based on global sequence alignment. We demonstrate that our algorithm is much faster and produces better results than the previously presented genetic algorithm. We also compare alphabets constructed on the base of universal substitution matrices like BLOSUM with our alphabets built through sequence alignment and propose a new coefficient describing the value of alphabets in the sequence alignment context. Finally we show that our simplified alphabets give better results in a sequence classification (using k-NN classifier), than most previously presented simplified alphabets and better than full 20-letter alphabet.

Keywords: amino acid alphabet, sequence alignment, substitution matrices, protein classification.

1 Introduction

Proteins are represented by 20 symbols, usually each symbol is a letter of the English alphabet and corresponds to one amino acid. However, in some cases, a simplified alphabet can be more convenient. In a simplified alphabet there are less than 20 symbols - each symbol from the simplified alphabet replaces a set of symbols from the original alphabet. In general, a simplified alphabet is built by grouping sets of symbols from a starting alphabet together - each group is represented by a new symbol. In this paper we consider if a simplified alphabet, which is as universal as possible, is the best choice for sequence alignment and for classifying a new protein. We propose a fast genetic algorithm for finding simplified alphabets and compare it to previously presented methods, where standard substitution matrices like BLOSUM are used. Next, we present a method of creating simplified alphabets with no substitution matrices based on the presented

* The research has been partially supported by grant No 3 T11C 002 29 received from Polish Ministry of Education and Science.

E. Marchiori, J.H. Moore, and J.C. Rajapakse (Eds.): EvoBIO 2007, LNCS 4447, pp. 122–131, 2007.

genetic algorithm and sequence alignment. We compare simplified alphabets using a new proper alignment coefficient. We also show that simplified alphabets can improve the correctness of amino acid sequence classification with respect to the full alphabet.

2 Previous Work

Simplified alphabets are used in several fields of bioinformatics. In sequence alignment (for protein classification), using simplified alphabets can provide more reliable results (see [1]) because it is often possible to change one amino acid to another in a protein sequence without any changes of the protein's properties. There are also some works ([3], [6], [11]) pertaining to how many amino acids are needed to fold a protein properly and whether all 20 amino acids are necessary to build all proteins that occur in Nature. There is a previously presented genetic algorithm proposed in [10] and a branch and bound algorithm designed in [2]. We compare produced alphabets to those based on the residue pair counts for the MJ and BLOSUM matrices ([7]), and to those based on the correlation coefficient ([9]).

3 Methods

It has been shown (see [2]) that alphabet simplifying can be related to the problem of set partitioning. The purpose is to divide an original set into disjoint subsets which completely cover the initial set. For example we can join any two symbols together and replace them by one new symbol. Starting from an original 20-letter amino acid alphabet we can repeat this operation as long as the alphabet has at least two symbols. It has also been shown that the problem of set partitioning is a hard computational problem because of the number of possible ways to create k subsets from a set of n elements can be calculated recursively following this expression:

$$S(n,k) = k * S(n-1,k) + S(n-1,k-1), 2 <= k <= n-1 \qquad (1)$$

where S(n,1) = S(n,n) = 1.

For a 20-letter alphabet and any count of subsets, there are more than $51\text{x}10^{12}$ possible simplified alphabets. When we consider, for example, only 8-letter simplified alphabets, there are more than $15\text{x}10^{12}$ such alphabets. The question is, which alphabet is the best and whether such an optimal alphabet exists.

3.1 Simple Rating Schema

One of the simplest ways to rate a reduced alphabet is to use one of the popular substitution matrices, like BLOSUM. In [2] a method has been proposed based on counting the total score for an alphabet. When we choose a substitution