Simplifying amino acid alphabets by means of a branch and bound algorithm and substitution matrices

N. Cannata, S. Toppo, C. Romualdi, G. Valle

CRIBI Biotechnology Centre, Università di Padova, via Ugo Bassi 58/B, 35131 Padova, Italy.

Protein and DNA are generally represented by sequences of letters. In a number of circumstances simplified alphabets, in which one or more letters are represented by the same symbol, have proved their potential utility in several fields of bioinformatics including searching for patterns occurring at an unexpected rate, studying protein folding and finding consensus sequences in multiple alignments. The main issue addressed here is the possibility of finding a general approach that would allow an exhaustive analysis of all the possible simplified alphabets, using substitution matrices like PAM and BLOSUM as a measure for scoring. Even user-defined matrices could be used, representing particular chemical-physical proprieties of the amino acid.

We found that very little work was done in this field. Some investigation was done using alphabets built with empirical approaches but a systematic way of exploring the space of all possible simplified alphabets was never undertaken.

Our computational approach has led to a computer program called AlphaSimp (Alphabet Simplifier) that can perform an exhaustive analysis of the possible simplified amino acid alphabets, using a branch and bound algorithm together with standard or user-defined substitution matrices. The program returns a ranked list of the highest-scoring simplified alphabets. When the extent of the simplification is limited and the simplified alphabets are maintained above ten symbols the program is able to complete the analysis in minutes or even seconds on a personal computer. However, the performance becomes worse, taking up to several hours, for highly simplified alphabets.

AlphaSimp and other accessory programs are available at http://bioinformatics.cribi.unipd.it/alphasimp.