**Abstract**

**Motivation and results:** An algorithm is described for the quick identification and evaluation of amino acid substitutions in multiple protein sequence alignment. The strategy is based on the calculation of a relative conservation index for an amino acid at each position of the alignment. The algorithm is implemented in the computer program POLINA (Protein Oriented LINear Analysis) which provides a summary of analysis in a format suitable for import into a graphing program.

**Availability:** A copy of source code is available upon request from the authors or can be downloaded via the WWW at http://www.geocities.com/Athens/4654/POLINA.html.

**Contact:** slevin@aecom.yu.edu; bsatir@aecom.yu.edu

The identification and evaluation of amino acid substitution in multiple sequence alignment are often difficult. It is impractical to search for substitutions visually and then use objective criteria for evaluating their importance (e.g. Dayhoff, 1978). A number of methods have been developed to aid in the interpretation of multiple sequence alignments, but most of them are geared toward defining regions which are functionally important (Devreux et al., 1984) or highly conserved (Livingstone and Barton, 1993). In this paper, we present a simple program to point out and quantitatively evaluate non-conservative amino acid replacements in multiple alignments of homologous proteins.

The method assumes that a number of necessary tasks have already been performed. First, a multiple sequence alignment must already exist. Secondly, there must exist a pre-defined measure of the similarity between two amino acids $a$ and $b$, which will be denoted $s(a, b)$. Good examples include the values constructed by Dayhoff (1978) or Jones et al. (1992), which measure the frequency of amino acid replacements or deletions in homologous protein pairs, or the values of Overlington et al. (1992) which are derived from structural information.

Assessment of the importance of a single amino acid substitution in position $r$ of protein $p$ is performed in three steps.

1. Evaluate $C(p, r)$, the conservation of residue $r$ in protein $p$, with respect to all amino acids in position $r$ of the alignment:

$$C(p, r) = \frac{\sum_{i=1}^{n} s(a_{p, r}, a_{i, r})}{n}$$

where $a_{p, r}$ denotes the amino acid in protein $p$ at residue $r$.

2. Evaluate $C^{all}(r)$, the conservation of residue $r$ among all proteins in the alignment:

$$C^{all}(r) = \frac{\sum_{i=1}^{n} C(i, r)}{n}$$

3. Construct a conservation index $I(p, r)$:

$$I(p, r) = \frac{C(p, r)}{C^{all}(r)}$$

Qualitatively, this method not only considers how conservative or statistically frequent a substitution is, but also compares it to the variability of amino acids in this position in all proteins in the alignment.

This algorithm was implemented in the computer program POLINA (Protein Oriented LInear Analysis). POLINA was written in C on a UNIX platform and is easily portable to other operating systems. A Java version of the program is being tested.

The program reads multiple alignment of an amino acid sequence of up to 100 proteins and can either analyze conservation of all residues in one protein or can analyze conservation of all residues of all proteins in the alignment. The output of the program is a file in column format that can be imported into a graphing program. An external file with an amino acid similarity table can be supplied by the user.

We have tested the program on an alignment of hemoglobin α-chains from different species (Dickerson and Geis, 1983). The program was able to detect and predict the importance of the point mutations in human hemoglobin α-chains (Figure 1).

It is important to understand that with more proteins in the alignment the program will produce better prediction values, although adding a large number of almost identical proteins will yield false-positive results.
Fig. 1. Relative conservation of normal (---) and variant (-----) human hemoglobin α-chains in an alignment of 12 hemoglobin α-chains from different mammalian species as analyzed using POLINA. Substitution from Gly to Arg at a variable position 15 (a) in hemoglobin variant Siam (Pootrakul et al., 1974) has little effect on hemoglobin function. However, substitution of a conserved Leu to Pro at position 48 (b) in variant Bab-Saadoun (Molchanova et al., 1992) affects the stability of the hemoglobin molecule. The relative frequency of amino acid replacement [as observed in 1572 examples of closely related proteins (Dayhoff, 1978)] was used as a similarity measure in this example. The program can also accept standard similarity matrices like PAM250 (with negative values for dissimilar amino acids) or custom similarity measures that are specific for the analyzed protein superfamily.

Unlike other methods (Livingstone and Barton, 1993), this program is made specifically to point out non-conservative mutations in large protein superfamilies. POLINA will be useful for verifying strong conservation regions for homology modeling and especially for finding mutations responsible for the change or loss of function in a member of a protein superfamily.

References


