Swaps in protein Sequences

Amit Fliess and Ron Unger
Faculty of Life Science, Bar-Ilan University, Israel
ron@biocom1.ls.biu.ac.il

To what extent proteins underwent swaps of fragments? Such events might be swaps of short fragments, swaps of structural and functional motifs, or fusion in different order of multiple domain proteins. This question is important both for theoretical understanding of the evolution of proteins and also has practical implications for protein sequence comparison. In order to analyze the question systematically, we are conducting a large scale survey of possible swaps and permutations among all pairs of protein from the Swissprot database.

We define a swap as a specific kind of sequence mutation in which a part of the sequence is being placed out of its original sequential order. for example, we define the pair of sequences: xxAxxBxCx and xxxBxAxx as a swap where A,B,C represent sequence fragments.

Identifying such swaps is not possible using standard sequence comparison packages [Russell and Ponting, 98]. Global sequence alignments, due to their linear nature, can not detect swaps at all. Local sequence matching procedures produce a list of high matching fragments, which need to be further processed in order to identify swaps. The main problem in the analysis stems from the fact that many sequences contain repeats, which may cause false positive pairs. For example, ABCB may look like a permutation of ABC, since the suffix of the first sequence CB is a swap of the suffix of the second sequence BC, but a more likely scenario is that B, the last fragment of the first sequence is a mere duplication.

In order to systematically study swaps and their properties the following scheme was designed: After eliminating highly homologous sequences, all pairs between 21,500 proteins from the Swissprot database were screened. These pairs were compared using WU-Blast 2 [Altschul et. al. 97], which produces, for each pair of proteins, a list of fragments that have strong local similarity. Automatic screening was used to find the pairs that may contain swaps. Filtering procedures were used to eliminate some (but not all) of the pairs that are related to duplications or have other irregularities. After few steps of filtering, a set of about 6500 candidate pairs of permutations was found. These pairs were visually inspected to find actual cases of swaps. This procedure left about 2000 cases. Preliminary analysis of these pairs reveals the following: Most cases were found in proteins longer than 400 residues. In almost all cases only one swap is found between a pair of proteins. In about 1500 of the cases, the swapped fragments are short relative to the similar fragments between the sequences, i.e. Ab to bA where b is much smaller then A. A typical size of the small swapped fragments is about 20 to 40 residues. In about 500 cases the swapped fragments are of the same size of the similar fragments. About 15-20% of the pairs involve proteins that are very similar, e.g. have identical function in different species. Several interesting examples will be shown. The theoretical and practical implications of these findings will be discussed.

References:

1. Russell RB, Ponting CP, [1998]. Protein fold irregularities that hinder sequence analysis. Curr Opin Struct Biol 8:364-371.

2. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ, [1997]. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389-3402.