Common Spatial Arrangements of Backbone Fragments in Proteins.

Nickolai N. Alexandrov, Katsutoshi Takahashi and Nobuhiro Go

Department of Chemistry, Faculty of Science, Kyoto University
Kyoto 606, Japan.

The number of the determined three-dimensional structures of proteins is growing steadily and at the present time about 300 distinct protein structures are known. This volume of data leads to considerable interest in their comparative studies. Several approaches have been developed for comparison of protein three-dimensional structures. All these previous methods compare three-dimensional structures that are continuous or have the same order in amino acid sequence. In this communication we report our initial results of search for common spatial arrangements of backbone fragments in proteins. In similar arrangements of fragments in a pair of proteins, fragments may also appear in different order in their respective amino acid sequences. This new type of three-dimensional structural similarity appears to be important for understanding the evolution of proteins, the mechanism of protein activity, and possible rules of protein folding.

Before making searches for similar spatial arrangements of backbone fragments we chose a measure of similarity and estimated the statistical significance of the similarities. The most natural measure of the distance between a pair of fragments is the root-mean-square distance (r.m.s.d.) between corresponding atoms in the mutually bestfitted position. In this communication we focus attention on the
backbone structures of polypeptide fragments and use the r.m.s.d. between Cα atoms. At first we study the distribution of the values of r.m.s.d. for pairs of fragments with no significant similarities. From this distribution we determine its lower bound RL as a function of the length of the fragments. This value is used to judge the significance of the similarity between a given pair of fragments with the r.m.s.d. value R. If R is significantly smaller than RL, then the pair is judged to be significantly similar. Thus, S = RL - R is defined as the similarity score. In this expression the linear approximation for RL is used in our communication.

We have developed a computer program SARF to search for common Spatial ARRangements of backbone Fragments (referred to as SARFs hereinafter) in a pair of proteins. A SARF is described by a set of Cα-atoms from one protein. This set is not necessarily continuous along the amino acid sequence. A common SARF is a pair of SARFs with a similarity score larger than the threshold value SC. Corresponding Cα atoms in the pair of SARFs may appear in a different order along the amino acid sequence in the compared proteins.

The program follows essentially the following algorithm. We start by finding similar small continuous fragments. For this purpose, the two protein structures are divided into overlapping fragments with a fixed length (usually six or seven residue long). All the possible pairs of fragments from that two proteins are superposed by the method of McLachlan, and we retain and compile a list of all pairs with the similarity score larger than the initial similarity score S0 (usually 0.8). Let the i-th pair in this list defined by a pair of fragments, Fm from one protein and Fn from the other, be denoted by P_i = P(F_m,F_n). Each pair P_i has an already calculated similarity score S_i. Then, to proceed to find common SARFs we temporarily unite two pairs. To do this, two pairs, say P_i = P(F_m,F_n) and P_j = P(F_k,F_l), are taken from the list of pairs and united fragments (candidates for SARFs) F_{mk} = F_m ∪ F_k and F_{nl} = F_n ∪ F_l are considered. Here we treat the united fragments, each of which is a set of Cα atoms, as ordered sets so that corresponding atoms appear in the same order in
the sets. As a consequence of this treatment Cα atoms may not appear in the natural order of the amino acid sequence in the united fragments. Then, the similarity score S_{ij} is calculated for a pair P_{ij} = \text{P(F_{mk},F_{nl})}. This calculation is done for all possible pairs i and j. After that, we find a pair P_{ij} with the largest similarity score amongst those satisfying the inequalities S_{ij}>S_{i} and S_{ij}>S_{j}. When such a pair is identified, we unite, now permanently, the pair i and j into one to create a new pair of similar SARFs and update the list of pairs. We repeat this procedure until no unification of SARFs will increase the similarity score. SARFs with the similarity score larger than the threshold value S_C are reported. In the current version of the program we set an upper limit on the size of SARFs equal to 48 residues. The actual program contains modifications to the algorithm to improve the the memory space requirement and the speed of calculation.

The above method is applied to detect common spatial arrangements between endothiapepsin (chosen arbitrarily) and all other proteins in the Protein Data Bank (PDB). Endothiapepsin, consisting of 330 amino acid residues, belongs to the family of acid proteases. It consists of two similar domains, formed by residues (-2)-159 and 162-326 (in PDB numeration), respectively.

Among 514 structures from PDB we have found 31 with the similarity score more than 3.00. This similarity score corresponds to r.m.s.d. = 3.00 Å for the 45-residue fragments. Eight of them are with proteins in the same family of acid proteases, i.e., pepsin, chymosin, penicillopepsin and rhizopuspepsin. Five cases are the already reported similarities with retrovirus proteases. It has been suggested that the virus proteases are active in dimer form and endothiapepsin's structure is a result of gene duplication and fusion. These similarities with the obvious evolutionary reason have the similarity score near 4.0-5.0.

The next similarity between endothiopepsin and carbonic anhydrase involves five β-strands, four of which form a well developed sheet, and one short α-helix. It is
interesting to mention that the order of the fragments in the amino acid sequence is not the same in these two SARFs. In both cases the active sites are located very near the β-sheet, but on opposite sides. Even though the similarity, involving as many as 42 residues is quite striking, it is difficult to understand its meaning without further investigations.

The common SARFs from endothiopepsin and prealbumin involve one β-sheet consisting of six strands. This β-sheet is formed by fragments from different domains in endothiopepsin and from different chains in prealbumin. This may imply an important common role of these SARFs in interaction between domains or between different polypeptide chains. Moreover, in both molecules they form the base of a channel where the active site in endothiopepsin and the hormone binding site in prealbumin are located. This might indicate that the structures formed by the common SARFs have common function, for example, to orient substrates appropriately to the active sites.

To summarize the results we first note that common SARFs are found not only with proteins in the same family, but also quite frequently with proteins with no obvious relations. Comparison of endothiapepsin with other proteins in PDB revealed cases of possibly evolutionary, functional and structural meaning. Note also that different parts of endothiapepsin constitute different SARFs that are common with different proteins. It appears that common SARFs exist rather frequently in proteins for a variety of reasons. Further studies are necessary to fully understand the biological meaning of variety of common SARFs.