A library of low-variance cores for protein families

Russ B. Altman*, Robert Schmidt*, & Mark Gerstein^

*Section on Medical Informatics
SUMC, MSOB X215
Stanford, CA 94305-5479
{altman, rschmidt}@camis.stanford.edu

^Department of Structural Biology
Fairchild Building, D109
Stanford, CA 94305
mbg@hyper.stanford.edu

http://www-camis.stanford.edu/projects/cores.html

As the number of known three-dimensional protein structures increases, it becomes possible to classify related structures into families. A number of such classifications have been reported, and many of these are now available over the internet [1,2]. These resources typically provide a multiple alignment based on structural superpositions that provide information about equivalent amino acid positions for the members of each family. Although some protein families only have one or a few members, many have sufficient members to allow a quantitative statistical analysis of the shared (and divergent) structural features.

We have previously reported a method for calculating a low variance core for a family of similar, aligned proteins [3,4]. The method defines those atoms in a multiple alignment that have very low spatial variance, and therefore assume essentially the same position in all members of the family. We have applied this method to the globin and immunoglobulin families, and found that the cores, as we define them, have great biological significance. For example, the cores have been related to folding intermediates, intron/exon organization, the identification of functional sites, and possible evolutionary relationships between families. In addition, we believe that these cores may have value for homology modeling and threading applications, because they offer very accurate positions for a subset of atoms after alignment. For these reasons, we have embarked upon the creation of a library of protein cores, generated automatically from the available protein family classifications. The resulting core library is being cross-indexed with the available classifications, to assist users in accessing cores of interest.

In this poster, we summarize (1) the number of protein cores generated, along with differences observed between different structural alignment techniques, (2) the statistics of protein cores (% of total protein in core, % of aligned residues in core, % of secondary structures in core) as we have computed them, (3) the format we have developed for reporting cores, including the full rank order of atom variability from most to least variable, and (4) the WWW interface we have developed for accessing the core library.

[1] Overington JP; Zhu ZY; Sali A; Johnson MS; Sowdhamini R; Louie GV; Blundell TL. Molecular recognition in protein families: a database of aligned three-dimensional structures of related proteins. Biochemical Society Transactions, 1993 Aug, 21 (Pt 3)(3): p. 597-604.

[2] Holm L; Sander C. The FSSP database of structurally aligned protein fold families. Nucleic Acids Research, 1994 Sep, 22(17): p. 3600-9.

[3] Altman, RB; Gerstein, M. Finding an average core structure: Application to the globins. Proceedings of Second International Conference on Intelligent Systems for Molecular Biology, Stanford, August 1994, AAAI Press, Menlo Park, p. 19-27.

[4] Gerstein, M; Altman, RB, Average Core Structures and Variability Measures for Protein Families: Application Immunoglobulins. Journal of Molecular Biology, 251, p. 161-175.

Figure 1. (TOP) The eight members of the aspartic proteinase family are shown at far left. The ellipsoids showing the volume of spatial variation for atoms that are determined to be noncore in the middle (green), and the ellipsoids showing the volume of spatial variation for atoms with low variance that are determined to be core on the right (red). (MIDDLE) The five members of the flavodoxin family drawn in the same manner. (BOTTOM) The eight members of the lipocalin family also displayed in the same manner.