Structural Neighbors and Structural Alignments: The Science Behind Entrez/3D

Stephen H. Bryant and Christopher W.V. Hogue

Presented at the IUCr Macromolecular Crystallography Computing School, August, 1996

Computational Biology Branch, National Center for Biotechnology Information, National Institutes of Health, 8600 Rockville Pike, Bethesda, MD 20894 USA

Entrez is an Internet tool for retrieval of information on the structure and function of biological macromolecules [1,2] (http://www4.ncbi.nlm.nih.gov/Entrez/). It provides daily-updated databases of molecular sequences, three-dimensional structures, and the Medline citations pertaining to molecular genetics. With simple term matching queries one may easily retrieve information on a molecule of interest from any of these sources. One may also easily "link" between databases, to retrieve, for example, the Medline citations contained within a molecular sequence or structure report.

Entrez's most powerful source of information on molecular structure and function, however, is provided by its "neighbor" database. The neighbors of a sequence are its homologs, as identified by a significant similarity score using the BLAST algorithm[3]. The neighbors of a Medline citation are articles which use surprisingly similar terms in their title and abstract[4]. Since biological functions are often conserved among members of a homology group, and/or described in the associated Medline abstracts, one may easily explore the structure-function relationships of an entire protein family by traversing these neighbor relationships.

Structural neighbor information in Entrez is based on a direct comparison of 3D structure. All of the roughly 10,000 domain substructures within the current Protein Data Bank[5] have been compared to one another using the VAST algorithm[6,7], and the structure- structure alignments and superpositions recorded. The VAST algorithm, for "Vector Alignment Search Tool", places great emphasis on the definition of the threshold of significant structural similarity. By focusing on similarities that are surprising in the statistical sense, one does not waste time examining many similarities of small substructures that occur by chance in protein structure comparison. Very many of the remaining similarities are examples of remote homology, often undetectable by sequence comparison. As such they may provide a broader view of the structure, function and evolution of a protein family.

At the heart of VAST's significance calculation is definition of the "unit" of tertiary structure similarity as pairs of secondary structure elements (SSE's) that have similar type, relative orientation, and connectivity. In comparing two protein domains the most surprising substructure similarity is that where

the sum of superposition scores across these "units" is greatest. The likelihood that this similarity would be seen by chance is then given as a simple product: the probability that one would obtain this score in drawing so many "units" at random, times the number of alternative SSE-pair combinations possible in the domain comparison, from which one has chosen the best. In practice one finds that the VAST significance threshold identifies similarities that span a sizable fraction of the structures compared, and it would appear that this theory corresponds to the subjective criteria long employed by crystallographers.

In addition to a listing of similar structures, neighbors within the Entrez 3D structure database contain detailed residue-by-residue alignments and transformation matrices for structural superposition. Alternative alignments are examined using a Gibbs sampling algorithm, beginning from the "seed" SSE-pair alignment. The optimal alignment is defined as that which is most surprising relative to the background distribution of alpha-carbon superposition residuals one obtains by chance drawing structural fragments at random. This definition provides an objective criterion with which to balance the well-known trade-off of lower superposition residuals versus more aligned residues. In practice refined alignments from VAST appear conservative, choosing a highly similar "core" substructure. In this superposition one easily identifies regions where protein evolution has modified the structure.

Structural neighbor calculations for Entrez are based on the MMDB database [2, 8] (http://www.ncbi.nlm.nih.gov/Structure/), a validated version of the Protein Data Bank in a computer-friendly form suitable for comparative analysis. Structural neighbors are presented via 3D molecular graphic images, using the Cn3D viewer that is distributed as part of the Entrez client software. Cn3D operates on a variety of computer platforms, including MacIntosh, Windows and Unix, and it provides a variety of algorithmic rendering schemes suitable for visualization of structural superpositions. Structure superposition data may also be easily exported from Entrez, most simply by writing PDB-format files rotated to the reference frame of a neighbor. In this way Entrez may serve as a starting point for detailed comparative analysis by structural biologists using other software to examine the patterns of structural conservation and change within a protein family.

References:

1. Schuler, G.D., Epstein, J.A., Ohkawa, H. and Kans J.A. (1996) Entrez: Molecular biology database and retrieval system. *Methods Enzymol.* 266, 141-162.
2. Hogue C.W.V., Ohkawa H., and Bryant, S.H. (1996) A dynamic look at structures: WWW-Entrez and the molecular modeling database. *Trends Biochem Sci. 1996*, 21, 226-229
3. Altshul, S.F., Gish W., Miller W., Myers E.W., Lipman D.J. (1990) Basic local alignment search tool. J. Mol. Biol, 215, 403-410. .
4. Wilbur, W.J., Yang Y.(1996) An analysis of statistical term strength and its use in the indexing and retrieval of molecular biology texts. *Comput. Biol. Med. 1996*, 26:209-222.
5. Abola, E.E., Bernstein, F.C., Bryant, S.H., Koetzle, T.F., Weng, J.C. (1987) Protein data bank. In Crys- tallographic databases: information content, software systems, scientific applications. Edited by Allen FH, Bergerhoff, G, Sievers R. Bonn, Chester, Cambridge: *International Union of Crystallography* 107-132.
6. Madej, T., Gibrat, J-F., and Bryant, S.H. (1995) Threading a database of protein cores. *Protein*

Struct. Funct. Genet. 23 356-369.
7. Gibrat, J-F., Madej, T., Bryant, S.H. (1996) Surprising similarities in structure comparison. *Current Opinion in Structural Biology*. 6, 377-385.
8. Ohkawa, H., Ostell, J., Bryant, S. (1995) MMDB: An ASN.1 specification for macromolecular structure. *ISMB* 3, 259-267.

Rev. 17 Feb 1997
