Assignment of structural domains in proteins: why is it so difficult?

Stella Veretnik[1], Ilya N. Shindyalov[1]. Phillip E. Bourne, [1,2]

Keywords: 3D protein structure, detection of 3D protein domains, automatic domain assignment methods, consensus approach, curated domain resources.

1 Introduction.

Structural domains are often considered to be basic units of protein structure. Assignment of structural domains from atomic coordinates is crucial for understanding protein evolution and function. Currently there is no good agreement among different assignment methods for what constitute the basic structural unit, underscoring the complexity of structural domain assignment. This work discusses tendencies of individual methods and highlights the problematic areas in assignment of structural domains by experts as well as by fully automated methods.

2 Methods.

Domain assignments were analyzed for three automatic methods (DALI[1], DomainParser[2], PDP[3]) and three expert methods (AUTHORS[4], CATH[5], SCOP[6]), using a 467-chains dataset assigned by all 6 methods. The following features were investigated: agreement on the number of assigned domains, agreement on domain boundaries, distribution of domain sizes and tendency toward assignment of discontinuous domains. Consensuses among automatic, expert and all methods were defined and used during comparison to tease out the behaviors specific to individual assignment methods or groups of methods.

3 Results and Discussion.

We observe that unambiguous domain assignments (when all methods agree on domain assignment) are confined predominantly to one-domain chains. Agreements among all methods in multi-domain chains are infrequent; in all cases the domains are compact and clearly spatially separated. For the majority of multi-domain proteins, there is no agreement on domain assignment among all methods. From the consensus analysis we observe that the majority of the difficulties of fully automated methods stem from overwhelming reliance on the structural cues (compactness/contact density) during domain assignments and the lack of functional/evolutionary information. Thus the cases in which domains are positioned close together are difficult or impossible for automatic methods to resolve. On the other hand, the differences in expert methods arise from different philosophical approaches underlying the specific methods. Authors of the structures (AUTHORS method) tend to define domains based on functionality, which may produce small and structurally not clearly defined domains. The creators of SCOP, on the other hand, often look for the largest common structure (fold) as a domain, which often consists of several distinctive structural units. The CATH method appears to strike a balance between sometimes contradictory structural, functional and evolutionary information. The inconsistencies in expert assignments are well reflected in the propensities of different fully automated methods, as those are trained and validated using a specific expert method, thus reflecting its philosophical biases. Detailed analysis

[1] San Diego Supercomputer Center, University of California, San Diego, 9500 Gilman Dr. La Jolla 92093-0537

[2] Department of Pharmacology, University of California, San Diego, 9500 Gilman Dr. La Jolla 92093

of structures which do not have consensus between the assignment methods regarding the number of assigned domains indicates the following problematic areas: (1) assignment of small domains, (2) discontinuous domains and unassigned regions in the structure, (3) splitting of the secondary structure elements between domains (if required), (4) convoluted domain interfaces and complicated architectures. Comprehensive domain re-definition, which takes into account the above issues is overdue and will be a great step toward improvement of domain definitions in multi-domain proteins, which represent (by an estimation [7]) 66-75% of the sequence database. Also, the intensive growth of 3D protein data demands fully automated approaches to be used to maintain currency and uniformity of domain information relative to the PDB.

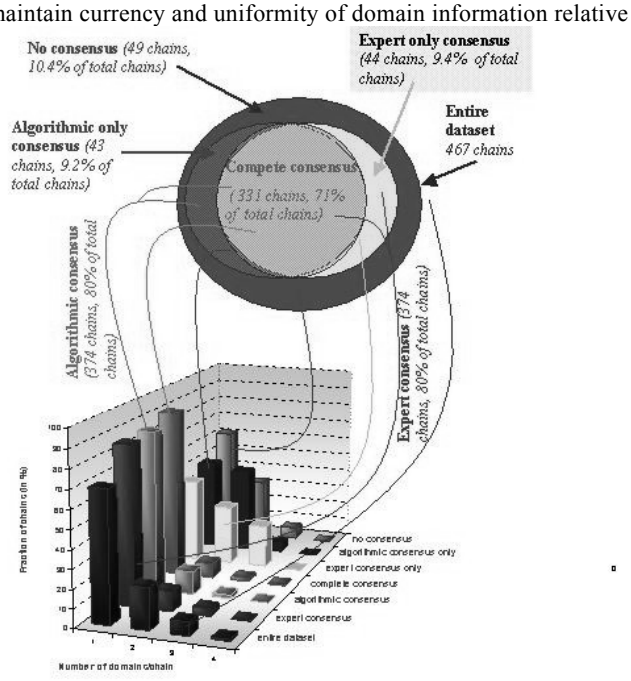

Figure1. Distribution of single- and multi-domain structures within different consensuses of domain assignment methods.

References

[1] Holm L., S. C. 1996 Mapping the protein universe. *Science* 273, 595-602.

[2] Guo, J-T. Xu, D. Kim, D. Xu, Y. 2003 Improving the Performance of DomainParser for Structural Domain Partition Using Neural Network, *Nucleic Acids Res*.31(3), 944-952.

[3]. Alexandrov, N. & Shindyalov, I. 2003 .PDP: protein domain parser. *Bioinformatics* 19, 429-430.

[4] Islam, S. A., Luo, J. & Sternberg, M. J. 1995 Identification and analysis of domains in proteins. *Protein Eng* 8, 513-25.

[5] Orengo, C. A., Michie, A. D., Jones, S., Jones, D. T., Swindells, M. B. & Thornton, J. M. 1997 CATH—a hierarchic classification of protein domain structures. *Structure* 5, 1093-108.

[6] Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. 1995. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol* 247, 536-40.

[7] Chothia C, Gough J, Vogel C, Teichmann SA. 2003 Evolution of the protein reprertoire. Science 300, 1701-03.