

Domain-based Protein Hierarchy and Detection of Semantically Significant Domain Architectures

Chin-Jen Ku,¹ Golan Yona,²

Keywords: domain-based protein hierarchy, multi-domain proteins

1 Introduction.

Regrouping biologically related proteins in an organized structure plays an important role in the study of the biological function of protein sequences. Such a structure is typically established by clustering proteins via some measures of similarity using *sequence-based* or *domain-based* methods. While the sequence-based approaches analyze the full-length protein sequences and assign those with high similarity scores to the same family (*e.g.* [4, 6]), the domain-based approaches decompose each protein into a sequence of units called *domains* which are believed to be crucial in the characterization of its biological properties. These domains are then grouped into domain families (*e.g.* [1, 3]) and the proteins are clustered based on their domain architecture [2]. Here we introduce two hierarchies of multi-domain proteins and investigate the biological relationships among proteins that are associated to hierarchically connected domain families. Moreover, we attempt to identify domain families that effectively characterize proteins with specific properties and propose to do so by searching for semantically significant domain families. We formalize the notion of semantic acquisition for a domain family and propose the means to detect it.

2 Domain-based hierarchy.

- **Q-hierarchy and S-hierarchy.** The Q-hierarchy is based on the *sequence* order among the domain families (*i.e.* sequences of domains). A sequence is said to be a *subsequence* of another one if the elements of the first sequence can be extracted one by one from the second while preserving the ordering. The S-hierarchy uses the more restrictive *string* order where a sequence is a *substring* of another if it is entirely subsumed in the second sequence without extraneous elements. Thus each hierarchy dictates a specific set of relationships between proteins with similar domain architectures.

- **Study of hierarchical relationships.** We distinguish the *parent-child* and the *sibling* relationships. Two domain families form a *parent-child* (PC) pair if one (parent) is a subsequence/substring of the other (child). On the other hand, two domain families are called *siblings* if their domain architectures are made of the same domain units, possibly in different ordering. Study of these relationships is interesting since they can reveal the effects of domain *deletion/insertion* and *permutation* on the biological properties of the proteins.

- **Validation of hierarchical relationships.** To study and evaluate the relationships defined above we use the Pfam database [1]. We assess the biological similarity between proteins based on the pairwise *semantic similarity* (ss) score derived from the Gene Ontology annotations [5]. For the PC relationship, the experiments show that the “distance” between a pair of PC domain families is strongly correlated with the difference in the number of domains L . As L grows, the proteins associated with the child family are increasingly more

¹Dep. of Computer Science, Cornell University, Ithaca, NY, USA. E-mail: kucj@ece.cornell.edu

²Dep. of Computer Science, Cornell University, Ithaca, NY, USA. E-mail: golan@cs.cornell.edu

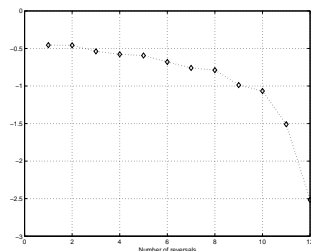


Figure 1: Difference $\Delta(D_1, D_2)$ vs. N_{rev} over the Q-hierarchy sibling domain families.

similar to each other compared to those associated with the parent family. This result verifies the validity of the hierarchical organization. It also highlights the fact that the domain units are carriers of biological information and additional domains in the child domain family increases the information content pertaining to the description of the child-related proteins. In addition, we observe that the siblings of a given domain family are not equally similar. Figure 1 shows that the average difference of ss scores $\Delta(D_1, D_2)$ between two groups of proteins associated to sibling domain families diminishes monotonically as the number of domain reversals N_{rev} grows, which indicates an increasing biological dissimilarity.

3 Semantic acquisition.

We address the issue of finding “signature” domain families that characterize proteins with specific properties. We reformulate this problem as identifying domain families that possess a *semantic meaning*. A domain family is said to acquire a semantic meaning (*i.e.* the semantic acquisition occurs) when it imposes a significant constraint on the structure of the adjacent domains in the protein for such a constraint limits the possible choices of domain extension and therefore its biological properties. To identify the occurrence of semantic acquisition, we propose a statistical procedure that analyzes the effect of appending an extra domain unit to a given family on the probability distribution of its *domain extension*. Our procedure has successfully identified several instances of semantic acquisition. These cases were verified through the 3D structure of the proteins in question.

References

- [1] Bateman A., Birney E., Cerruti L., Durbin R., Etwiller L., Eddy S.R., Griffiths-Jones S., Howe K.L., Marshall M. and Sonnhammer E.L.L. (2002), “The Pfam protein families database,” *Nucl. A. R.*, vol. 30, pp. 276-280.
- [2] Coin L., Bateman A. and Durbin R. (2004), “Enhanced protein domain discovery using taxonomy,” *BMC Informatics*, vol. 5, no. 56.
- [3] Corpet F., Servant F., Gouzy J. and Kahn D. (2000), “ProDom and ProDom-CG: tools for protein domain analysis and whole genome comparisons,” *Nucl. A. R.*, vol. 28, pp. 267-269.
- [4] Tatusov R.L., Galperin M.Y., Natale D.A. and Koonin E.V. (2000), “The COGS database: a tool for genome-scale analysis of protein functions and evolution,” *Nucl. A. R.*, v. 28, pp. 33-36.
- [5] The Gene Ontology Consortium, (2000) “Gene Ontology: tool for the unification of biology,” *Nature Genetics*, vol. 25, pp. 25-29.
- [6] Yona G., Linial N. and Linial M. (2000), “ProtoMap: automatic classification of protein sequences and hierarchy of protein families,” *Nucl. A. R.*, vol. 28, no. 1, pp. 49-55.