Organization and Prediction of Structural Domains in Multidomain Proteins

J. Gorodkin[1,2], O. Lund[1]*, and S. Brunak[1]
[1]Center for Biological Sequence Analysis, Department of Biotechnology,
The Technical University of Denmark, Building 208, DK–2800 Lyngby, Denmark
[2]Department of Ecology and Genetics, The Institute of Biological Sciences,
University of Aarhus, Building 540, Ny Munkegade, DK-8000 Aarhus C, Denmark.
Emails: gorodkin@pop.bio.aau.dk, olund@strubix.dk, brunak@cbs.dtu.dk

Abstract

Structural domains in multidomain proteins that have known three-dimensional structure were investigated. Eukaryotic and eubacterial domains were found to follow different length distributions. The length distribution of eukaryotic domains could be fitted well by a gamma distribution that is similar to that proposed for primordial proteins by White [1]. As the length distribution of eubacterial domains is fitted well by a log-normal distribution, the results support the random-origin hypothesis [1] as well as an introns-late exon-shuffling scheme. Eukaryotic multidomain proteins are found to have a preference for having a domain boundary located in the middle of the chain regardless of the chain length. This is in contrast to eubacterial multidomain proteins for which there are no such preference. In agreement with this, eubacterial domains in the same multidomain protein are found more often to be of different sizes than what is the case for eukaryotic domains. It is concluded that eubacterial domain organization is more complex than the organization of eukaryotic domains. This suggests that eubactarial domains more often are discontinuous in sequence and that eukaryotic domains more often are continuous in sequence. These findings are related to folding mechanisms and domain duplication. In agreement with this we find that multidomain proteins with continuous domain organization have "heavy" boundaries (K, E preference) and that those with discontinuous organization have "light" boundaries (G, S, D preference). The two boundary types are in addition proline rich and contain a characteristic coil dominated secondary structure motif. The information about domain boundary composition and domain organization is used to construct neural network schemes that predicts the location of domain boundaries. A server will be made available at www.cbs.dtu.dk.

References

[1] S. H. White. The evolution of proteins from random amino acid sequences: II. evidence from the statistical distributions of the lengths of modern protein sequences. *J. Mol. Evol.*, 38:383–394, 1994.

[2] J. Gorodkin, O. Lund, and S. Brunak. In prep., 1999.

*Present address: Structural Bioinformatics Advanced Technologies A/S, Agern Alle 3, DK-2970 Hørsholm, Denmark.