J.C. Nacher[1]
e-mail: nacher@fun.ac.jp
M. Hayashida[2]
e-mail: morihiro@kuicr.kyoto-u.ac.jp
T. Akutsu[2]
e-mail: takutsu@kuicr.kyoto-u.ac.jp
[1]Department of Complex and Intelligent Systems, Future University Hakodate, Japan
[2]Bioinformatics Center, Institute for Chemical Research, Kyoto University, Uji, Japan

Data analysis and mathematical modeling of internal duplication process in multi-domain proteins

Multi-domain proteins have likely been shaped by selective genome growth dynamics during evolution. Emergence of new protein domains allows to perform new functions as well as to create polypeptide structures that fold on a biologically feasible time scale. Although the dynamics of genome growth through shuffling of protein domains have been studied extensively over decades, recent experimental observations of a significantly large number of domain repeats of several domains from the same family suggests that one more process involving domain recombination may still remain hidden [1, 2]. Here we examine the protein domain statistics retrieved from Pfam, SMART, Gene3D, ProDom and TIGRFAMs databases and consisting of 68 eukaryotic, 56 archaeal, and 929 bacterial organisms. We show that this analysis confirms earlier observations [3] and extends them to numerous organisms in the three kingdoms of life. The results show that the number of total protein domains and the number of domain families in a protein are governed by different statistical laws. While the former follows a power-law distribution, the latter exhibits an exponential statistics. We develop a methodology and propose an evolutionary dynamics model, based on a rate equation formalism, and consisting of domain fusion, mutation, protein duplication and internal duplication processes. We then demonstrate that these distinct distributions are in fact rooted in the internal domain duplication mechanism. The analytical results derived from the evolutionary dynamics model as well as computer simulation show that this domain-repeats event generates a wide number of domains in a protein while at the same time preserving a thin number of domain families across proteome species. To our knowledge, this is the first mathematical model of protein domain evolution that explicitly takes into account the effect of internal duplication mechanism and provides analytical solution. These findings bring in our view new insights into the fundamental mechanisms governing genome expansion with potential implications in the development of protein interaction network models and related evolutionary studies.

References

[1] A.D. Moore, *Arrangements in the modular evolution of proteins* Trends in Biochemical Sciences **33** 444-451.
[2] A.K. Björklund, D. Ekman and A. Elofsson, *Expansion of protein domain repeats* PLoS Computational Biology **2**, e114.

[3] E.V. Koonin, Y.I. Wolf and G. P. Karev, *The structure of protein universe and genome evolution* Nature **420**, 218-223.