10. False occurrences of functional motifs on protein sequences highlight evolutionary constraints.

Allegra Via 1, Federico Gherardini1, Enrico Ferraro 1, Gianpaolo Scalia Tomba 2, Gabriele Ausiello 1 and Manuela Helmer-Citterich 1

1Centro di Bioinformatica Molecolare, Department of Biology, University of Rome Tor Vergata, Roma
2Department of Mathematics, University of Rome Tor Vergata, Roma

We carried out a statistical and biological analysis of the relationship between the false predictions of sequence functional patterns and their occurrences in random sequences. Results are discussed in the perspective of the evolutionary mechanisms that might have affected the chance occurrence of functional motifs in biological sequences.

A functional motif is a set of residues that is characteristic of a specific biochemical function. The detection of a functional motif in yet uncharacterized protein sequences is a well-established method for assigning function to proteins. A critical problem, however, concerns the evaluation of the false prediction rate of a motif in sequence databases, i.e. the significance of finding a motif in several proteins. The number of false positive (FP) matches of a pattern has been often assessed from the number of its occurrences expected by chance (E) for the mere aggregation of letters in a database search, as can be calculated from the residues frequency in the database ([1], [2], [3], [4]) . The relationship between E (expected) and FP (observed), however, has not been thoroughly investigated so far. It is reasonable to expect that the function fitting the set of data (E,FP) is linear, but it is not clear a priori if there are exceptions (i.e. number of false predictions on a biological database sensitively greater or lower than the expected number of hits on the corresponding random database), how frequent they are, and the reason why they occur. In this work, we carried out a statistical study of such relationship and an analysis of the unexpected behaviours, thus providing insights into the random nature of protein sequences. The analysis described in this work was performed on 1226 PROSITE patterns in the form of regular expression and based on three sequence datasets: the complete Swiss-Prot database (sprot100), the set of H. sapiens sequences, and the set of S. cerevisiae sequences derived from sprot100. We assumed for the expectation E of a pattern P, the mean number of hits on N database randomizations. In order to preserve the local sequence composition the datasets were randomized by reshuffling each single sequence. As a control, the statistical analysis was also performed on PROSITE reversed patterns, which represent a reliable sample of non-functional patterns. Our results show that a) the relationship (E, FP) is linear, b) the great majority of functional motifs (group II) have a number of false occurrences comparable to the number of matches on a random database, c) there is a group (group I) of PROSITE patterns for which FP >> E and another one (group III) for which E >> FP. Both groups I and II are outside the 95% confidence interval around the value E = FP. A detailed analysis of patterns belonging to each group revealed several interesting features. In particular patterns belonging to different groups do share specific statistical and biological properties. We compared groups using the patterns information content ([1], [2]) – as a statistical parameter – and the tendency of their false positive hits of being in either disordered or ordered/globular regions of proteins, as a biological parameter.
Our findings suggest diverse fascinating mechanisms and constraints occurring during evolution, which might “regulate” the random appearance of functional motifs in protein sequences.

[1] Jonassen, I., Collins, J.F. & Higgins, D.G. (1995) Finding flexible patterns in unaligned protein sequences. Protein Sci. 4, 1587-95.

[2] Jonassen, I., Eidhammer, I., Grindhaug, S.H. & Taylor, W.R. (2000) Searching the protein structure databank with weak sequence patterns and structural constraints. J.Mol.Biol. 304, 599-619.

[3] Nevill-Manning, C.G., Wu, T.D. & Brutlag, D.L. (1998) Highly specific protein sequence motifs for genome analysis. Proc. Nat. Acad. Sci. USA 95, 5865-5871

[4] Sternberg, M.J.E. (1991) Library of common protein motifs. Nature 349, 111.