Electronic Supplementary Material (ESI) for Molecular BioSystems

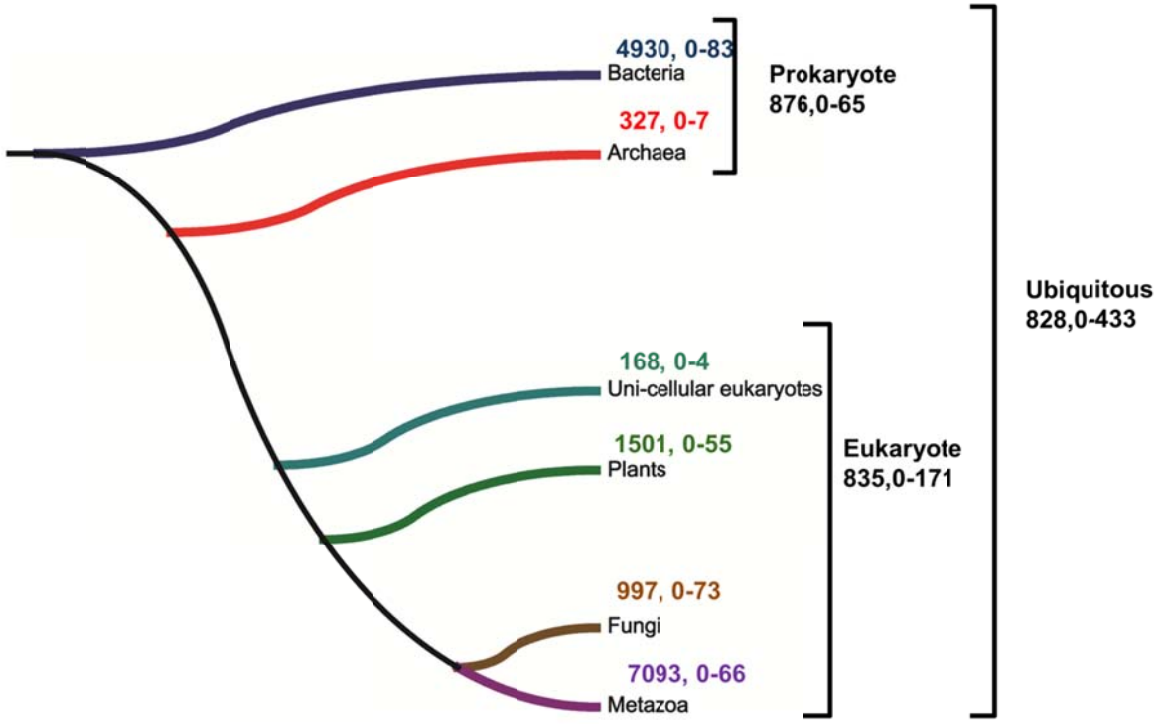
Supplementary Information accompanying with the manuscript titled:

"Tethering preferences of domain families co-occurring in multi-domain proteins"

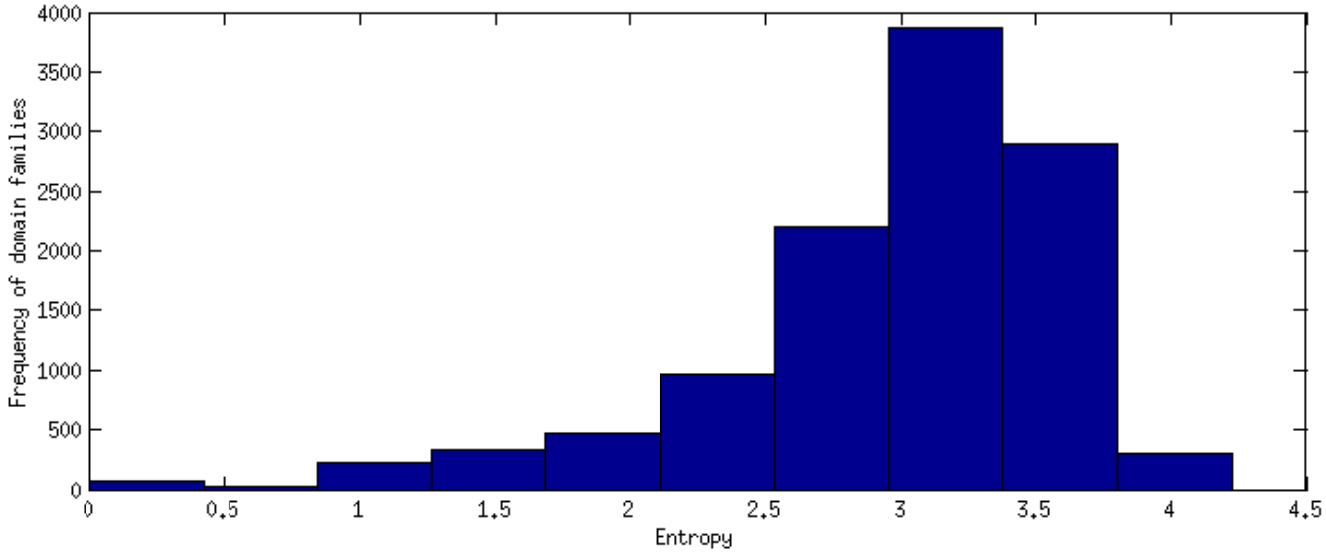
Smita Mohanty, Mansi Purvar, Naryanswamy Srinivasan* and Rekha Nambudiry

***corresponding author: ns@mbu.iisc.ernet.in**

Supplementary Figures (SF1-3)



Supplementary Figure SF1 Collapsed species tree showing the distribution of domain families found in individual groups and the tethering number range exhibited by these domain families. The entire species tree is based on 31 orthologues proteins found in 191 species [1] and they cluster into six major groups of organisms. In each group the number of domain families uniquely found and their tethering number range has been indicated. (tree generated using MEGA [2])

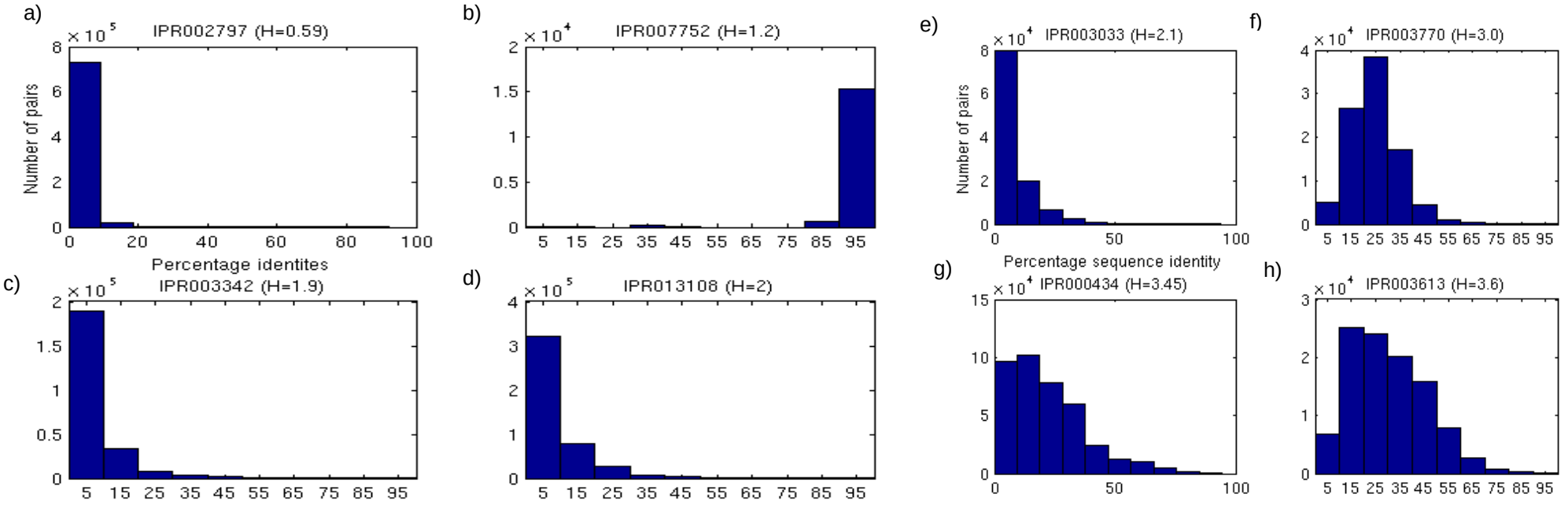


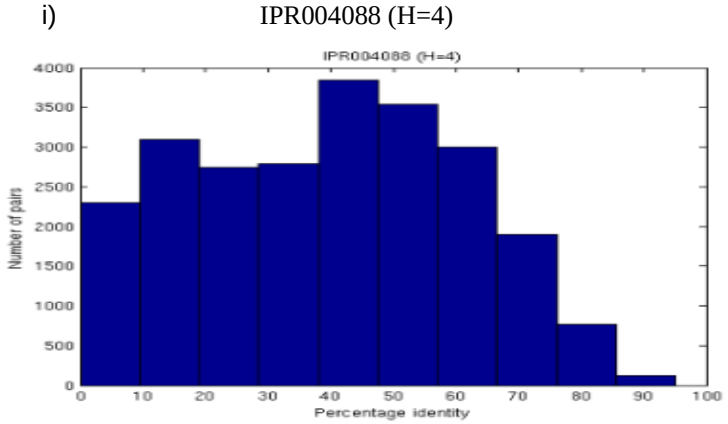
Supplementary Figure SF2 Distribution of sequence entropy values for the domain families with more than 10 members. Histogram shows the distribution of frequency of domain families with

respect to entropy values. A large fraction of domain families have entropy values from 3-3.5, while there are very few examples of domain families with very high or low entropy values.

Supplementary Text 1 Distribution of sequence entropy across the entire dataset

Nearly 50% of the domain families have entropy values in the range 3-3.6, indicating a spread in the sequence identity values within these families (average entropy for the dataset is 3.1 ±0.5). There are few domain families showing very low entropy (317 domain families have entropy of 2.1 or lower) or very high entropy values (19 domain families with entropy values of 4 or higher). From the distribution of entropy value for the entire dataset it can be seen that the tendency of a family to have a highly homogeneous population or completely heterogeneous population is low. Most domain families show wide distribution of pairwise sequence identities and thus contain both sequences with high and low sequence identity at varying frequencies. Supplementary figure SF3 shows the distribution of the pairwise sequence identity obtained for a range of families exhibiting different entropy values. Domain families with entropy values nearing zero (Supplementary figure SF3 a-d) can be of three kinds a) poorly conserved (Supplementary figure SF3a) b) highly conserved (Supplementary figure SF3b) c) moderately conserved (Supplementary figure SF3c, SF3d), depending on the bin having the maximum frequency. Typically such examples are few in the dataset (317 domain families have an entropy of 2.1 or lower) and in most of these cases it is observed that a large proportion of the sequence pairs have poor identity between them. One of the reasons for such a skew in the distribution could be because of the lack of completely sequenced genomes, as a result of which several sequences have not been accounted for. However, there are also examples of domain families where such diversity may have certain biological advantage. For example, the PAS domain family (IPR000014, Entropy=1.9) is found in many signaling proteins across all the three kingdoms. They function as signal sensors and hence presence of enormous sequence diversity will enable this domain and the protein containing it to respond to a wide array of stimuli. Supplementary figure SF3e-h shows examples of domain families with entropy values ranging from 3 to 3.6. A large proportion of domain families in the dataset lie within this range. From the graphs it can be observed that the identity across sequences within families varies across a much wider range rather than peaking at a single bin. At the higher entropy values the distribution becomes more uniform (Figure SF3i). Families showing moderate to high entropy value (3 onwards) contain mixed population of sequences leading to sequence variation, such that the dispersion increases as the entropy value tends to maximum entropy effectively helps in capturing the extent of variation amongst sequences within families.





Supplementary Figure SF3 Sequence identity distributions for few domain families at different entropy values. Each histogram above shows the distribution of percentage identities obtained upon comparing all possible pairs of sequences within a domain family. The X-axis depicts the percentage identity and the Y-axis represents the number of pairs. As mentioned in the text the entropy values are obtained by grouping all the pairwise identities into 20 bins. The above histogram shows the frequency of occurrence of pairs in each of these bins for few domain families. At lower entropy value it can be seen that there are singular peaks indicating homogeneous families (a-c) as the entropy values increase the peaks broaden thus tending towards more heterogeneous population of sequences (d-f) finally tending towards completely heterogeneous family where all bins are nearly equally populated (g).

References

1. F. D. Ciccarelli, T. Doerks, C. von Mering, C. J. Creevey, B. Snel and P. Bork, *Science*, 2006, **311**, 1283-1287.
2. K. Tamura, D. Peterson, N. Peterson, G. Stecher, M. Nei and S. Kumar, *Mol Biol Evol*, **28**, 2731-2739.