BMC Systems Biology

Oral presentation

Linking evolution of protein structures through fragments

Sanne Abeln* and Charlotte M Deane

Address: Department of Statistics, University of Oxford, 1 South Parks Road, Oxford, OX1 3TG, UK

Email: Sanne Abeln* - abeln@stats.ox.ac.uk

* Corresponding author

from BioSysBio 2007: Systems Biology, Bioinformatics and Synthetic Biology
Manchester, UK. 11–13 January 2007

Published: 8 May 2007

BMC Systems Biology 2007, **1**(Suppl 1):S12 doi:10.1186/1752-0509-1-S1-S12

This abstract is available from: http://www.biomedcentral.com/1752-0509/1?issue=S1

Motivation

At present there is no universal understanding of how proteins can change topology during evolution, and how such pathways can be determined in a systematic way. The ability to create links between fold topologies would have important consequences for structural classification, structure prediction and homology modeling. Several methods based on geometrical measures have been proposed to create links between topologies, e.g. [1,2]. It has proven difficult, however, to show the evolutionary relevance of such links. Here we use our previously developed age measure for protein superfamilies [3] to investigate the relationship between structural fragments and protein structure evolution.

Results and discussion

We used a set of pairwise fragments to create a network of structural links between superfamilies. In total 1.2e-8, 1.5e-7, 2.7e-6 and 1.1e-5 fragments were generated of lengths 10, 15, 20 and 30 respectively. When comparing the number of fragment-links that young and old superfamilies make with other superfamilies, it becomes clear that the distribution of younger folds is skewed towards fewer links (Figure 1). Similarly we can compare the number of links that each superfamily has with a set of young and a set of old superfamilies. Again most superfamilies share significantly fewer links with the group of young superfamilies (Figure 2). New proteins are thought to be created through duplication and point mutations of structural domains. Here we show (the first) evidence that this might also occur on a scale below the domain level: fragments are shared more often with older superfamilies, which is expected in a model where new topologies can be built through an assembly of, or multiple insertions of, fragments from existing proteins. A little care has to be taken here as these results could also be caused by a scenario of convergent evolution, which would drive the inclusion of more stable fragments. However, the differences between age groups become stronger, with increased fragment length (Figure 2). When increasing the fragment length the probability of convergence should decrease contradicting the above argument. These results have important implications for structure prediction, as it may explain why current 'fragment based' modelling approaches are so successful.

Methods

Fragments

The fragment library generated for this study, contains fragment-pairs of length 10, 15, 20 and 30, with a maximum allowed gap-lengths of 2, 3, 4 and 6 respectively. All fragments are based on pairwise comparisons between structural domain as defined by SCOP. The pairs are scored for similarity purely on structural grounds, using the coordinates of the c-alpha atoms. This is to avoid bias, based on sequence similarity. All possible pairwise fragments between two domains of the given lengths are first screened and aligned using a method similar to the pre-filter used by MAMMOTH [4]. Each fragment pair with an alignment score above a threshold is then superimposed giving the c-alpha RMSD score for the fragment pair.

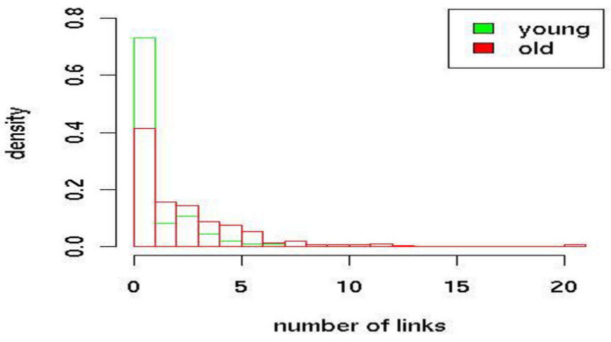

Figure 1
Density chart showing the distribution of links between superfamilies based on fragment-pairs of length 30. The distributions for 'young' and 'old' superfamilies are shown separately, with younger fold having significantly fewer links (Wilcoxon unpaired test: p-value = 1.2e-09). Note that the distribution of links per superfamily is not normally distributed.

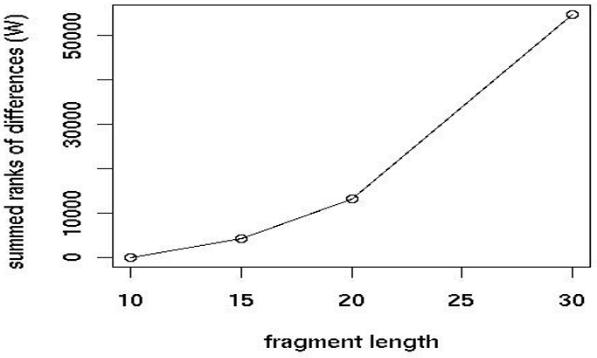

Figure 2
Fragment length versus W score of Wilcoxon's signed-rank test. Wilcoxon singed-rank tests was performed on data for each fragment length: for each superfamily the number of links it makes with a set of young superfamilies and a set of old superfamilies is compared. The values are normalised for the size of the age groups. Since the number of compared superfamilies in each test set are identical, the W scores can be compared directly.

Age estimates

Age estimates for protein folds or superfamilies are generated using fold recognition of structural domains on a set of completed genomes. The occurrence patterns of such predictions, are analysed with a parsimony algorithm to estimate an age for a superfamily, for more details see [3]. The age of a superfamily is based on a score between [0.0,1.0], with 1.0 indicating the superfamily was estimated to be present at the root of the species tree (oldest), and 0.0 estimating that the superfamily was created at the leaf level (youngest). Here an 'old' fold is defined as a fold with an age of 1.0, and a 'young' fold with an age < 0.5.

Linking Folds

Some fragments might be over-represented (e.g. secondary structure is not considered) therefore the number of shared fragments needs to be normalised for the number of times a fragment occurs. Friedberg and Godzik (2005) used a superfamily based normalisation to overcome this problem [2]. We use a similar approach, although the fragment-pairs in this study are based on structural similarity only. (whereas Friedberg and Godzik (2005) used a combination of sequence and structural similarity). A link between two superfamilies (I and J) is established when f(I, J) > 0.1, which is calculated as:

$$f(I,J) = \frac{Sim(I,J)}{min(Sim(A-I,I),Sim(A-J,J))} \text{ if } I \neq J$$

Here Sim(A, B) is the number of shared fragments between two set of domains (e.g. superfamilies), and A is the set of all domains. In this study we do not consider self-similarity of superfamilies.

Conclusion

We show that younger folds have relatively fewer shared fragments with other folds, than old protein folds. This may indicate that evolutionary links above superfamily or fold level could be established, through such shared fragments.

References

1. Taylor WR: **A 'periodic table' for protein structures.** *Nature* 2002, **416(6881):**657-660.
2. Friedberg I, Godzik A: **Fragnostic: walking through protein structure space.** *Nucleic Acids Res* 2005:W249-W251.
3. Winstanley HF, Abeln S, Deane CM: **How old is your fold?** *Bioinformatics* 2005, **21(Suppl 1):**i449-i458.
4. Ortiz AR, Strauss CE, Olmea O: **MAMMOTH (matching molecular models obtained from theory): an automated method for model comparison.** *Protein Sci* 2002, **11(11):**2606-2621.