Evolution of Multi-domain Proteins by Gene Fusion and Fission

Sarah K. Kummerfeld1*, Christine Vogel1, Martin Madera1, Mary Pacold1,2 and Sarah A. Teichmann1

[1]MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, England and
[2] Dept Computer Science, University of Illinois at Urbana, 1304 W. Springfield Ave., Urbana, IL 61801, USA.

Received line

ABSTRACT

During evolution genes may undergo recombination to produce more complex proteins by gene fusion. The reverse process generates multiple less complex proteins by fission. Considering proteins from 131 genomes, we established rates of gene fusion and fission. We found 2869 groups of multi-domain proteins that exist as single proteins in some organisms, and two or more smaller proteins in others. Applying maximum parsimony to species trees, we found fusion events were four times more common than fission. Analysing the functions of these proteins reveals that one third of split proteins are subunits of a single function complex, while two thirds have separate functions in the same pathway or as components of a multi-functional complex.

Contact: skk@mrc-lmb.cam.ac.uk

INTRODUCTION

The main mechanisms of gene evolution are duplication, recombination and sequence divergence. Recombination leads to rearrangement of domains, the evolutionary units of genes and proteins. Fusion and fission of genes are particular types of recombination, which result in one long composite protein in one organism, or two or more smaller split proteins in another organism.

An example of this is the multi-functional fatty acid synthase complex, which has just two polypeptide chains in yeast and mammals, while in *E. coli* six chains with separate activities form the same complex. In fact, several yeast and *E. coli* enzymes of small molecule metabolism that are orthologous have evolved by fusion or fission (Jardine et al. 2002). There have been several previous studies concerned with identifying such groups of proteins related by either fusion or fission for the purpose of predicting functional relatedness of proteins (Enright et al. 1999, Marcotte et al. 1999, Yanai et al. 2001, Enright & Ouzounis, 2001). Here, we study such groups of proteins in order to estimate the relative rates of fusion and fission in genomes.

In the only previous genome analysis known to us that distinguished between fusion and fission events, Snel et al. (2000) used sequence comparisons and generated 252 phylogenetic trees for groups of proteins in 17 prokaryotes. They compared the number of composite and split proteins, concluding that fusion occurs more often than fission. Our data set is an order of magnitude larger: 2869 trees with 131 prokaryotic, bacterial and eukaryotic genomes. With this large data set, we quantitatively determine the relatively rates of fusion and fission by identifying individual events across the tree of life.

EVOLUTIONARY RELATIONSHIPS BETWEEN PROTEINS

We studied proteins in terms of their domains in order to detect distant relationships across 131 genomes (17 Archaea, 98 Bacteria and 16 Eukarya). In the Structural Classification of Proteins (SCOP) database (Murzin et al. 1995), domains are defined as evolutionary units which can exist on their own in a protein, or shuffle independently. Evolutionarily related domains are part of the same superfamily. We used structural domain assignments for 131 completely sequenced genomes taken from the SUPERFAMILY (v. 1.63) database (Gough et al. 2001).

Assigning structural domains to protein sequences means that more distant orthologs can be detected. In order to perform a comprehensive analysis of highly divergent sequences in distantly related organisms, we consider proteins at the level of domain architecture. By domain architecture, we mean the sequential arrangement of domains within a protein as assigned by SUPERFAMILY. Sequences with identical domain architectures may have low sequence identity due to divergence. However, structural studies by Apic et al. 2001 and Bashton and Chothia 2002, as well as our own comparison of domain architectures with the INPARANOID database (Remm et al. 2001), show that proteins of a particular domain architecture are homologs in the sense of having evolved from a common ancestor. From the 638,712 sequences in the SUPERFAMILY database, we considered the 200,455 with complete domain assignments. This included 7116 distinct domain architectures.

IDENTIFYING COMPOSITE AND SPLIT PROTEINS

Within the set of 7116 domain architectures, we wanted to identify those that represented proteins which evolved by fusion or fission. We looked for domain architectures present as a single protein, the composite form, in at least one genome, and as a set of shorter proteins, the split form, which joined end-to-end have the same domain architecture.

These composite and split domain architectures represent orthologous proteins in different genomes, where the composite protein was split at some stage by fission, or the split proteins fused.

We consider each group of composite/split domain architectures to be candidates for having undergone fusion and/or fission events. Grouping domain architectures into sets of composite/split yielded a total of 2869 candidates.

TREES OF COMPOSITE AND SPLIT DOMAIN ARCHITECTURES

For each of the 2869 candidates, we labeled each genome as composite or split and then used the tree of life to determine whether these proteins most likely arose though gene fusion or fission. The rates of fusion and fission were the same given two different species trees: the standard phylogeny of the NCBI taxonomy (http://www.ncbi.nlm.nih.gov/Taxonomy/taxonomyhome.html/) and the taxonomy proposed by Wolf et al. (2002). The assumption of an underlying species tree has been used previously by Snel et al. (2002) and Mirkin et al. (2003) for the purpose of identifying duplications and horizontal transfer, but it has not been used for calculation of fusion and fission rates before.

By assuming a species tree, we circumvent the need to build phylogenetic trees from protein sequences. This means that very divergent proteins which do not give sufficient information to build a tree, can be considered at the level of domain architecture. The main drawback of our approach is that it assumes the species tree is correct. This assumption is wrong in two cases: if the proteins are not orthologous or if horizontal transfer has occurred amongst the group of proteins. We confirmed the orthology of proteins within each domain architecture group by comparison with the INPARNOID database assignments. To evaluate the impact of horizontal transfer, we considered each event in the 2869 trees, identifying those that may have arisen through horizontal transfer. This showed that only 4% of events to be candidates for horizontal transfer. Therefore, we can confidently use domain architecture groups to extract orthologous proteins from genomes.

IDENTIFYING FUSION VERSUS FISSION EVENTS

Given the set of 2869 composite/split groups, we wanted to determine the mechanism of evolution: fusion or fission. Based on maximum parsimony, the ancestors within the species tree were assigned as either split, composite or unknown. Those labeled as unknown were cases where the descendants had equal representation in composite and split and the immediate ancestor was also unknown. An example of a tree and the assignment process is shown in Figure 1.

Maximum parsimony was also used to identify the position of fusion/fission events within the tree. For each tree, we identified the point(s) of change between composite and split forms, assuming that fusion and fission are equally likely. A single event was counted for each ancestor that was different from one or more of its children. For example, if the ancestor were composite, and two of its children split, this would be counted as one fission event. The majority of trees had a single event.

FUSION IS FOUR TIMES MORE COMMON THAN FISSION

Using our maximum parsimony method, we establish the point(s) in each tree where a fusion or fission event occurred. Across all 2869 trees, the number of fusion events is roughly four-fold higher than the number of fission events. This was a consistent trend in all kingdoms (see Table 1). In 763 cases we could not determine whether the events were fusions or fissions, because the tree was perfectly balanced.

The predominance of fusion versus fission is in agreement with the genetic mechanisms that lead to these events.

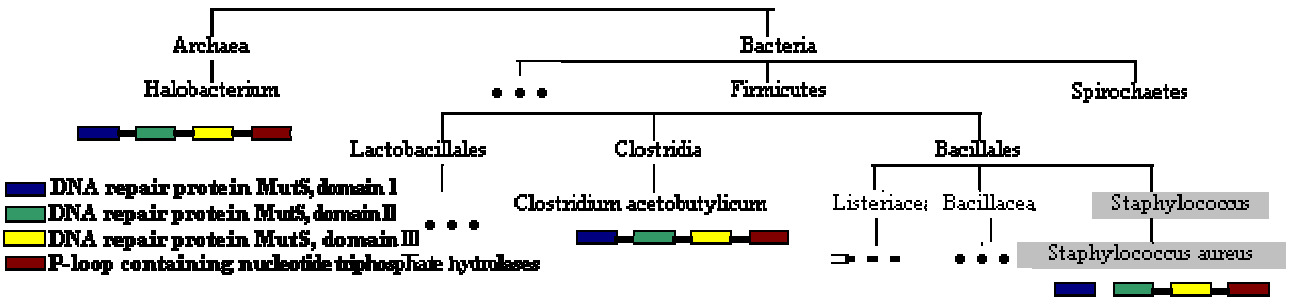

Figure 1. Identifying fusion or fission in a tree. This an example of one of the trees considered in our analysis. Only genomes that have this particular domain architecture (DNA repair protein MutS, domain I, domain II, domain III, P-loop) are shown. Sections of the tree have been omitted for simplicity, these are indicated by three dots. The leaves of the tree represent genomes. The composite or split form of the domain architectures is indicated by colour, white for composite and grey for split, and by the domain architectures below each genome. In this tree, there is a single genome with the split form of the architecture (*Staphylococcus aureus*), and all other genomes have the composite form. The ancestors have been allocated as split based on maximum parsimony, and thus the fission event takes place at the point in the tree between Bacillales where Staphylococcus branch off. The COGs annotation suggests that the proteins represented here are related to the DNA repair protein, MutS and probably exist as a heterodimer.

Fusion involves a loss of the terminal region of one gene and the initial regulatory regions of another, while fission involves the gain of these regulatory signals. Thus, fusion is simpler to achieve genetically, and has evidently produced more frequently retained protein products than fission events.

Table 1. Distribution of fusion and fission events across the three kingdoms

Kingdom	Fusion events	Fission events	Fusion/fissions
Eukarya	582	140	4.16
Archaea	294	58	5.07
Bacteria	1496	382	3.92

Table 2. Functional groupings of sets of proteins within each tree.

Functional group	#
Single COG	
Heteromultimeric enzymes	25
Two-component type signal transduction system proteins	8
Transcriptional regulation	4
Chaperones	3
Electron transfer chain proteins	2
Inteins	1
Unclassified	45
Total	88
Multiple COGs	
Components of multi-functional complex	25
Two-component type signal transduction system proteins	24
Consecutive enzymes in metabolic pathway	15
Consecutive components of electron transfer chain	4
ABC transporters	3
Unclassified	103
Total	174

FUNCTIONAL RELATIONSHIPS OF PROTEINS

We analysed the functional relationships of split proteins in a subset of 262 trees by using annotation from the Clusters of Orthologous Groups (COGs) database (Tatusov et al. 1997, Tatusov et al. 2001). Split proteins that are in the same COG are subunits of a protein with a single activity or function, like the DNA repair protein MutS. The majority of proteins in this set are heteromultimeric enzymes (25 out of 44 annotated groups of proteins), as shown in Table 2. Other proteins are two-component signal transduction proteins, transcriptional regulators, chaperones, electron transfer proteins and an intein. In cases where different COGs matched each the split domain architectures consistently across the tree, the split proteins have independent activities. Many of these are multi-functional proteins with several subunits, such as the DNA polymerase, or multi-functional complexes.

This functional analysis shows that the ratio of single to multi-functional proteins that have undergone fusion or fission is one (88) to two (174). Furthermore, it is evident that at least half the relationships between split proteins are a physical interaction, as for all the proteins with a single function and the multi-functional complexes. Many other split proteins are part of two-component type signal transduction pathways, enzymes that are involved in consecutive steps in metabolic pathways, or electron transfer chain proteins.

CONCLUSION

This study shows that the dominant process for evolution of composite proteins and their split counterparts is the fusion of smaller components to form larger ones rather than the reverse process. Across all trees in our data set, there were four times more fusions than fissions. The high rate of fusion compared to fission shows that multi-domain proteins can evolve by fusion of small proteins, and that fission is relatively rare. Furthermore, the vast majority of proteins and domain architectures is involved in only one fusion or fission event. This suggests that most domain architectures have evolved only once, and proteins with identical domain architecture have descended from the same common ancestor.

By mapping Clusters of Orthologous Groups onto each of the trees, we were able to determine the functional relationships of the different components. In one third of the trees, the individual split proteins belong to one single COG. Thus they associate to carry out a single function, such as representing the two subunits of a heterodimer, or different domains of two-component signal transduction system proteins. The other two thirds of trees contain split proteins belonging to separate COGs, because they have independent functions. These proteins are part of multi-functional complexes or enzymes involved in consecutive steps of metabolic pathways for instance.

ACKNOWLEDGEMENTS

Thanks to Graeme Mitchison, Cyrus Chothia, Julian Gough.

REFERENCES

Apic, G. et al. 2001. *J. Mol. Biol.* **310**: 311--325.
Bashton, M., and Chothia, C. 2002. ***J. Mol. Biol.*** **315**: 927--939.
Enright A.J. 1999. ***Nature*** **402**: 86--90.
Enright A.J. 2001. ***Genome Biology.*** **2**: 00341--00347.
Gough, J. 2002. ***Nucleic Acids Res*** **30**: 268--272.
Jardine, O. 2002. ***Genome Res.*** **12**: 916--929.
Marcotte E.M. 1999. ***Science*** **285**: 751--753.
Mirkin, B.G. 2003. ***BMC Evol. Biol*** **3**: 2.
Murzin, A.G. 1995. ***J. Mol. Bio.*** **247**: 536--540.
Remm M. 2001. ***Journal of Molecular Biology*** **314**: 1041--1052.
Snel B. 2000. ***Trends Genet.*** **16**: 9--11.
Snel B. 2002. ***Genome Res.*** **12**: 17--25.
Tatusov, R.L. 1997. ***Science.*** **278**: 631--637.
Tatusov, R.L. 2001. ***Nucleic Acids Res.*** **29**: 22--28.
Wolf, Y.I. 2002. ***Trends Genet.*** **18**: 472--479.
Yanai I. 2001. *Proc. Natl. Acad. Sci. USA*. **98**: 7940--7945.