2

Genome evolution: gene fusion versus gene fission

Berend Snel, Peer Bork and Martijn A. Huynen

Trends in Genetics **16** (2000) 9-11

Introduction

With the advent of complete genome sequencing, it has become possible to study gene evolution on a genome-wide scale (for an overview of sequenced genomes see http://www.tigr.org). Here, we present a systematic analysis of two principal processes in molecular evolution: the fusion and fission of genes, events that have so far mainly been recognized and described in individual cases (Leffers *et al* 1989 and Zakharova *et al* 1999). We quantify fusion and fission of orthologous genes (Fitch 1970) in completely sequenced prokaryotic genomes. As fission and fusion events of orthologous genes are unlikely to reflect a change in their function, genome-wide, rather than gene-specific, trends can be observed. The estimates of the occurrence of gene fission and gene fusion that we obtain are subsequently compared with each other and across the various genomes.

Methods

To obtain a candidate set of orthologous genes that underwent fission or fusion, we began our analysis with Smith-Waterman sequence comparisons (Smith and Waterman 1981, Pearson 1998) of all open reading frames (ORFs) from 17 completely sequenced genomes (see Table 1). For each pair of genomes we determined pairs of genes with highest, significant ($e < 0.01$, where e is the expected number of false positives in homology detection), bidirectional levels of identity, which we considered potential orthologs. We allowed a gene from a genome A to have more than one ortholog in a genome B if the alignments of the genes of B with the gene of A did not overlap with each other (Huynen and Bork 1998), providing the candidates for fission and/or fusion. Subsequently, families of orthologous proteins of these candidates were collected from the genomes. To ensure that our families consisted only of orthologous genes, we used additional information from relative levels of similarities to other genes, conservation of gene order (synteny) and, if necessary, genes in species that were not originally included in the analysis (Huynen and Bork 1998). Phylogenetic trees of these families were made and the distribution of the different gene organizations, either present as separate genes or as one gene, was mapped to the respective leaves. Considering scenarios with one single protein, as well as two split proteins, as the ancestral state, we determined the explanation of the distribution of organizations over the tree that required the smallest number of fission and/or fusion events (see Fig. 2.1 for an example). In determining this, we took into account only the reliable parts of the tree (high bootstrap values) and constructed trees for the parts as well as for the complete protein.

In addition, we analysed the DNA sequence of adjacent split genes that are present in only one species. Using the frameshift program of the (http://shag.embl-heidelberg.de:8000/Bic/; http://www.cgen.com), we tested if those split genes underwent a fission event that was generated by a single nucleotide frameshift deletion or insertion. These fissions then are either frameshift sequencing errors, or result from recent frameshift mutations. For example, the resequencing of a region of the *Mycoplasma pneumoniae* genome that contains three ORFs encoding fragments of the R subunit of the restriction modification system, which were generated by frameshifts that we also detected, has shown that the split organization is the actual organization (Himmelreich *et*

al. 1997**)**. In general, one cannot distinguish between the two possibilities based only on sequence data. Therefore, we put those putative fissions in a separate category, hereafter referred to as `frameshift'. The fissions for which we are certain that they occurred as such, we refer to as `genuine'.

Table 2.1 Number of gene organisations resulting from fission and fusion

Species[b]	Genome Size[a]	Fusion	Fission		
			Total	Genuine	Frameshift
Mycoplasma genitalium	468	2	2	1	1
Mycoplasma pneumoniae	677	2	1	0	1
Rickettsia prowazekii	834	6	2	0	2
Borrelia burgdorferi	850	3	1	1	0
Chlamydia trachomatis	876	8	0	0	0
Treponema pallidum	1031	6	0	0	0
Aquifex aeolicus	1522	12	13	8	5
Helicobacter pylori 26695	1590	9	0	0	0
Haemophilus influenzae	1717	18	13	3	10
Methanococcus jannaschii	1735	12	7	5	2
Methanobacterium thermoautotrophicum	1871	16	18	5	13
Pyrococcus horikoshii	2061	4	3	3	0
Archeoglobus fulgidus	2407	19	9	8	1
Synechocystis PCC6803	3168	24	4	4	0
Mycobacterium tuberculosis	3924	36	4	1	3
Bacillus subtilis	4100	19	1	1	0
Escherichia coli	4290	33	10	2	8

[a]Genome size in number of predicted genes

[b]Thermophilic species are shown in bold

Results and discussion

Numerous cases of fusion and fission (see Table 2.1, Fig. 2.1, and http://www.bork.embl-heidelberg.de/~snel/genetable.txt) were found that allow us to sketch the major trends. (1) Fusion occurs more often than fission (Table 2.1). The prevalence of fusion can be expected because there is a benefit to fusion in that it allows for the physical coupling of functions that are biologically coupled (Marcotte *et al.*, 1999). The number of genes resulting from fusion increases with genome size, which is to be expected, because a larger pool of genes by chance contains a larger pool of fused genes. (2) Genuine gene

fission is mainly observed in *Aquifex aeolicus* and the four archaeal species. All these species are thermophiles, and they contain significantly more split genes resulting from a genuine fission than non-thermophiles ($p < 0.01$ using the Mann¯Whitney test, see Table 2.1). This suggests that, at high temperatures, there is an increase in mutations leading to split genes, because larger thermal fluctuations lead to an increased error rate in replication. Alternatively, split genes might reflect an adaptation to high temperatures. If we assume that the number of errors that occur in the process of creating a functional protein from DNA (e.g. errors in transcription, translation or folding) is proportional to the sequence length, then, with for example a 10% error rate per 300 base pairs, 81% of one protein of 200 amino acids will be functional (90% × 90%). However, when two separate proteins code for two units of 100 amino acids each, 90% of the proteins will be functional [(90% + 90%)/2]. At higher temperatures, the error rate increases owing to larger thermal fluctuations (Jaenicke and Boehm 1998), and therefore this difference becomes more important. The increased impact of this process will then result in an increased advantage of having separate subunits coding for a certain protein complex.

Frameshift fissions appear not to be restricted to a specific type of organism (Table 2.1), but some genomes contain considerably more than others. This could mean that these genomes might contain more sequencing errors. Alternatively, if these fissions are recent frameshift mutations that render genes biologically inactive, this might mean that in these organisms there is a reduced selection for the functionality of certain genes, because these strains live under rich and constant conditions (see Burns *et al* (1995) for an example).

Recently, Marcotte *et al.*(1999) showed that proteins with homologs fused together in one protein are likely to interact. However, this prediction method has a high proportion of false positives (82%). We observe that the vast majority of pairs of genes whose orthologs are fused are either part of the same complex, or function in the same pathway. Thus, by considering only orthologs, the fraction of false positives can be substantially decreased, albeit at a price of reducing the number of proteins to which the method applies.

No general pattern was found in the functions of the genes that underwent a fission event. Genes resulting from a fission event are often annotated as hypothetical, because the split forms a problem in annotation of function (Bork and Koonin 1998). The reverse, fusion proteins being annotated as having only one of two functions, has also been observed.

Here for the first time, we have systematically and comprehensively surveyed the occurrence of gene fission and fusion. We find a correlation of fission with thermophily and argue that this lifestyle results in a selective pressure for the split organization of genes. As such, it is an example of the relation between phenotype and its composing parts. Cross-level relations like this stand at the core of genome function and evolution, and we expect that our understanding of them will eventually allow us to elucidate the principles that govern the dynamics of genome evolution.

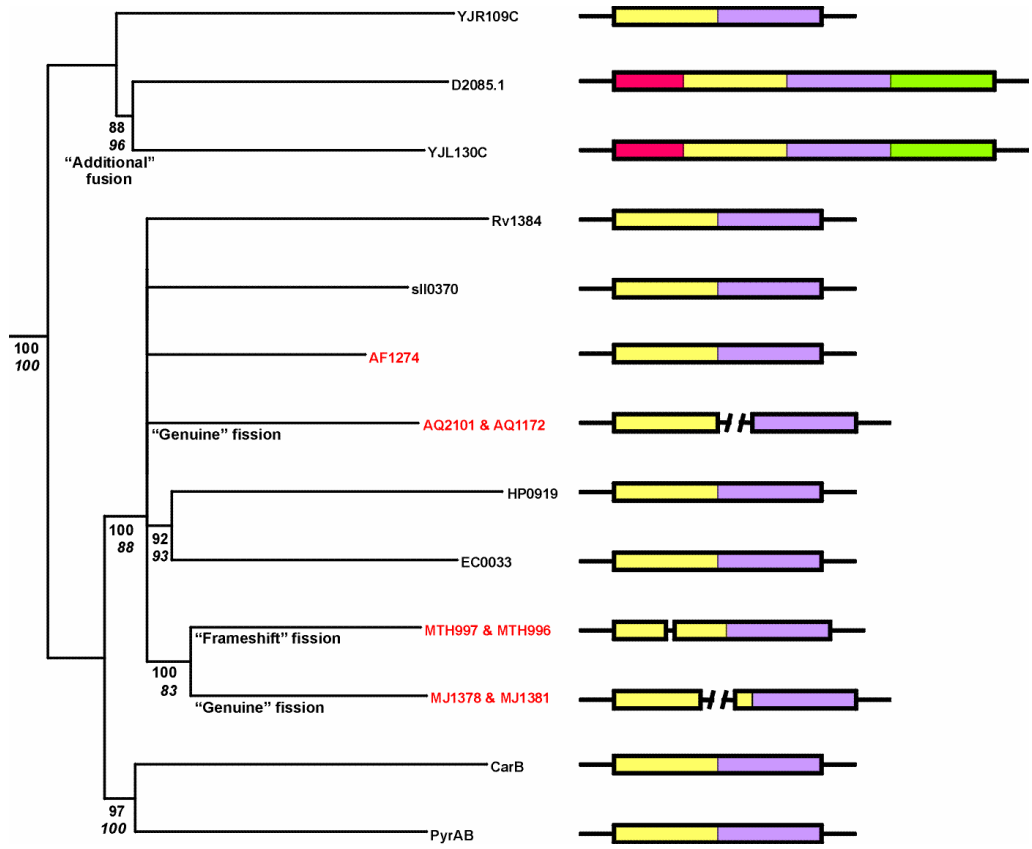

Figure 2.1. **The evolutionary history of carbamoyl phosphate synthase B (CarB).** The history of CarB contains fission events, and it illustrates some of the methodological challenges in determining fission and fusion. CarB is a large protein (900 amino acids) containing two major domains that are homologous to each other and that probably arose by an internal duplication. In *Aquifex aeolicus*, *Methanococcus jannaschii* and *Methanobacterium thermoautotrophicum*, CarB is encoded by two separate genes coding for different parts of the protein. In all the three cases, the split is in a different location, ruling out the possibility that all the fissions share a common origin, or that all of them still have the primitive state. The carB open reading frames in *M. thermoautotrophicum* are adjacent, and analysis of the DNA suggests that a frameshift insertion caused this organization. In *M. jannaschii* the location of the split, which is not between the two major domains, suggests that a fission event led to this organization. The split is not located in the structural domain involved in catalysis, but rather in a structural domain involved in oligomerization (Thoden *et al.* 1999); the enzyme is thus probably still active. The most parsimonious scenario is that the ancestral state of this family was one single protein, because then one fission in *M. jannaschii*, one fission in *A. aeolicus* and one frameshift mutation in *M. thermoautotrophicum* are sufficient to explain the present organization. By contrast, an ancestral state of two separate proteins requires seven fusions, one fission and one frameshift to explain the present-day situation. Phylogenetic trees were constructed from a multiple alignment of the complete CarB protein, and we used the internal duplication to root the tree of the complete protein (Iwabe *et al.* 1989). Shown is a schematic consensus tree from maximum likelihood (constructed using) (Strimmer and von Haeseler, 1996) and neighbour joining (constructed using) (Thompson *et al.* 1994) methods. Only clusters and bootstrap values with an average bootstrap value higher than 90% are shown. The numbers in normal case are the neighbour-joining bootstrap values, and those in italics are the reliability values. Fission events are shown under those branches where they occur in the most parsimonious scenario. The `additional' fusion of eukaryotic CarB with other domains is shown under its branch. At the leaves of the tree, a schematic drawing of the organization of the different carB genes is shown. The yellow box denotes the N-terminal domain of regular CarB, the purple box denotes the C-terminal of regular CarB, the red box is the CarA domain, and the green box is the aspartate transcarbamylase domain. YJR109C and YJL130C are from *Saccharomyces cerevisiae*, D2085.1 is from *Caenorhabditis elegans*, Rv1384 is from *Mycobacterium tuberculosis*, sll0370 from *Synechocystis* sp., AF1274 is from *Archaeoglobus fulgidus*, AQ2101 and AQ1172 are from *A. aeolicus*, HP0919 is from *Helicobacter pylori*, EC0033 is from *Escherichia coli*, MTH997 and MTH996 are from *M. thermoautotrophicum*, MJ1378 and MJ1381 are from *M. jannaschii*, and CarB and PyrAB are from *Bacillus subtilis*.