Gene fusion/fission is a major contributor to evolution of multi-domain bacterial proteins

Sophie Pasek¹,²,*, Jean-Loup Risler¹ and Pierre Brézellec¹

¹Laboratoire Statistique et Génome, 523 Place des Terrasses, 91034 Evry cedex, France and
²Solsusience, Biopôle Clermont-Limagne, 63360 Saint-Beauzire, France

Received on February 13, 2006; revised on March 22, 2006; accepted on April 3, 2006
Advance Access publication April 6, 2006
Associate Editor: Alex Bateman

ABSTRACT

Most proteins comprise one or several domains. New domain architectures can be created by combining previously existing domains. The elementary events that create new domain architectures may be categorized into three classes, namely domain(s) insertion or deletion (indel), exchange and repetition. Using 'DomainTeam', a tool dedicated to the search for microsynteny of domains, we quantified the relative contribution of these events. This tool allowed us to collect homologous bacterial genes encoding proteins that have obviously evolved by modular assembly of domains. We show that indels are the most frequent elementary events and that they occur in most cases at either the N- or C-terminus of the proteins. As revealed by the genomic neighbourhood/context of the corresponding genes, we show that a substantial number of these terminal indels are the consequence of gene fusions/fissions. We provide evidence showing that the contribution of gene fusion/fission to the evolution of multi-domain bacterial proteins is lower-bounded by 27% and upper-bounded by 64%. We conclude that gene fusion/fission is a major contributor to the evolution of multi-domain bacterial proteins.

Contact: pasek@genopole.cnrs.fr

Supplementary Information: Supplementary data are available at http://stat.genopole.cnrs.fr/domainteams/Bioinformatics/results.html

INTRODUCTION

Most of the proteins harbour two or more domains [such as those stored in SCOP (Andreeva et al., 2004) or Pfam (Bateman et al., 2004)], which results in a wide variety of domain combinations (Bornberg-Bauer et al., 2005; Orengo and Thornton, 2005). Since domains are considered as essential units for the modular assembly of new genes (Doolittle, 1995; Patthy, 2003; Vogel et al., 2004a), statistics on these combinations and on the distribution of the number of domains in proteins have been extensively analysed (Koonin et al., 2002; Vogel et al., 2004b). Recently, Björklund and collaborators (Björklund et al., 2005) have introduced a novel measure, called Domain Distance, which they define as the number of unmatched domains in an alignment of two domain architectures. Using this measure, they were able to quantify the elementary events [i.e. domain(s) insertion/deletion (indel), repetition and exchange] that distinguish a protein from its closest neighbour. However, to date, little is known about the relationships between these elementary events and the molecular mechanisms they originate from. We report here an analysis aiming at finding which molecular mechanisms are the sources of new domain combinations.

To investigate this question, we first searched for proteins that have obviously evolved by modular assembly of domains. The search for modular reshaped homologs, i.e. proteins encoded by genes derived from a common ancestor, is not as simple as it could seem [see Fitch (2000) and Koonin (2005)]. First, the impact of ‘evolutionary/elementary’ events on homology is that different parts (encoding distinct domains) of genes in one species may be orthologous to different genes in another species (in case of a gene fusion for instance). Second, classical methods based on sequence similarities cannot detect properly those homologous relatives that do not possess strictly the same domain architecture (Weiner et al., 2005). On the contrary, relying exclusively on the domain architectures to conclude on homology may result in linking too weakly related proteins. This can bias the quantification of the elementary events. As an example, consider two proteins p1 and p2 of respective domain architectures AC and ABC (where A, B and C are domains). One may infer that an internal insertion (deletion) of domain B occurred between p1 and p2. However, if p1 and p2 are weakly related and if there exists another protein p3 of domain architecture AB closer to p2, one would rather infer a terminal insertion (deletion) of domain C between p2 and p3. This is the reason why we searched only for strongly related proteins and based our search for homologs on the syntenic context of the genes. This one was determined using the DomainTeam software (Pasek et al., 2005, http://stat.genopole.cnrs.fr/domainteams/). In a first step, DomainTeam splits the proteins into their PfamA domains (Bateman et al., 2004). It then searches across several genomes for strings of domains that are conserved in their content but not necessarily in their order.

Using a definition of homology based on both domains and the syntenic context, we then collected sets of homologous proteins containing at least one reshaped protein, i.e. sets in which at least one protein differed from all the other proteins by one and only one elementary event. The subsequent analysis of these sets showed that (1) internal domain(s) indel and domain exchange are rare events whereas indels at either the N- or C-terminus are the most common events, (2) the genomic contexts of those genes reshaped by terminal indels reveal that a substantial number of them originate from gene fusion/fission. We show that the contribution of gene fusion/fission events to the evolution of multi-domain...
bacterial proteins is bounded between 27 and 64%. We conclude that gene fusion/fission is a major contributor to modular evolution of multi-domain bacterial proteins.

MATERIALS AND METHODS

Domain architecture definition

The domain architecture of a protein is defined as the ordered pattern of its PfamA domains (Bateman et al., 2004) from the N- to the C-terminus.

Definition of the elementary events

The elementary events that create new domain architectures can be categorized into three different classes: domain(s) exchange, indel and repetition. Exchange of domain is the substitution of one domain for another. Insertion (resp. deletion) is the addition (resp. excision) of a new domain(s) other than the adjacent domain(s) whereas repetitions are the addition of the same domain(s) as one of the adjacent domain(s). Domain(s) indel can be divided into two categories depending on the position of the indel. An internal indel is an indel which occurs in the middle of a protein (i.e. between two domains). A terminal indel occurs at either the N- or C-terminus of a protein.

Similarity between domain architectures

The similarity between two domain architectures \( Arch_1 \) and \( Arch_2 \) is defined as the ratio of intersection/cardinal where

1. intersection is defined as the number of domains that appear in both architectures and
2. cardinal is defined as \( \max(\text{card}_1, \text{card}_2) \) where \( \text{card}_1 \) (resp. \( \text{card}_2 \)) is the number of domains that compose \( Arch_1 \) (resp. \( Arch_2 \)).

Two identical domain architectures have a similarity value of 1 and, conversely, architectures with no domain in common have a similarity value of 0. Note that in this study, we imposed that (1) at least two domains have to be shared by the two architectures to consider that a similarity value can be calculated, (2) two domain architectures which differ by more than one elementary event are not taken into account.

Genomic context: syntenies of domains

The synteny context of the genes was determined using the DomainTeam software [see Pasek et al. (2005), http://stat.genopole.cnrs.fr/domainteams/]. In the first step, DomainTeam splits the proteins into their PfamA domains (Bateman et al., 2004). It then searches across several genomes for strings of domains that are conserved in their content but not necessarily in their order. A set of such ‘conserved strings’ is called a ‘domain team’ whereas each ‘conserved string’ is called an occurrence (Fig. 2).

DomainTeam is a tool that allows to process simultaneously intra-genomic and inter-genomic comparisons. The user-defined parameter \( \delta \), which specifies the maximal number of ‘foreign’ domains inserted...
between two domains belonging to the team, was set to 2. We discarded from this study all the domain teams having a score < 90 [see Pasek et al. (2005) for the definition of the score of a domain team].

**Identification of sets of homologous and reshaped proteins**

Homologous proteins (i.e. proteins encoded by genes deriving from a common ancestor) are defined as follows:

1. They are located in the same syntenic context (i.e. in two different occurrences of the same domain team).
2. Their domain architectures are the most similar in the domain team (where similarity is defined in the section ‘Similarity between domain architectures’).

A reshaped protein is defined as a protein which differs from its homolog(s) by one and only one elementary event.

Sets of homologous proteins containing at least one reshaped protein were built by considering each pair of occurrences in a domain team and by performing an all by all protein domain architecture comparison. For instance, in the example given in Figure 2, HI0147 is detected as a terminal indel with respect to its homolog VC1780.

The results have been manually verified by considering also the Pfam ‘context domains’ [Context domains are added by Pfam when a highly probable domain of a protein is not detected since its signature is lower than the PfamA threshold (Coin et al., 2003)] or the SMART domains (Letunic et al., 2004).

**Bacterial sets**

The bacterial sets used in this study are as follows:

- **Gram+**: Bacillus subtilis, Bifidobacterium longum, Clostridium perfringens, Corynebacterium efficiens, Deinococcus radiodurans, Enterococcus faecalis, Lactococcus lactis, Lactobacillus plantarum, Listeria monocytogenes, Mycobacterium leprae, Oceanobacillus iheyensis, Staphylococcus aureus N315, Streptococcus agalactiae serotype V.

The PfamA annotations pertaining to the above-mentioned proteomes were downloaded from ftp://ftp.sanger.ac.uk/pub/databases/Pfam/database-files

**RESULTS**

We ran ‘DomainTeam’ on two sets of complete bacterial genomes (see Materials and Methods). The first set comprised 15 Gram-negative bacteria and the second 13 Gram-positive (see Materials and Methods). Homologous reshaped proteins were searched for in the 8491 best-scoring domain teams (see Materials and Methods). We rejected those multi-domain proteins that could result from more than one ‘elementary event’, i.e. domain indel, exchange or repetition (see Fig. 1a and Materials and Methods).
Moreover, in order to fairly evaluate the relative proportion of the elementary events, we retained only the reshaped proteins with at least three domains. Otherwise, the position (internal or terminal) of the domains is irrelevant. Indeed, we observed that the vast majority of the two-domain reshaped proteins correspond to either N- or C-terminal indels. Considering these two-domain reshaped proteins would have led to underestimate internal indel. Finally, 141 sets of homologous proteins, each set containing at least one reshaped protein, were selected for analysis (see Supplementary Material Table S1 for the list of the 141 sets). These sets were classified according to the elementary events defined in Materials and Methods (Fig. 1a and b). Table 1 shows that the domain teams cover ~70% of the genes of the 28 bacteria considered in this study, providing strong support to the conclusions of our analysis.

The contribution of gene fusion/fission events to the evolution of bacterial multi-domain proteins is lower-bounded by 27% Indels are the most frequent events (95 out of 141, see Table 2). Among indels, the most numerous ones are terminal indels (90 out of 95, see Table 2), which substantiates a study carried out by Björklund and co-workers (Björklund et al., 2005). A statistical analysis shows that the number of terminal indels compared with internal indels is significantly greater than that expected by chance (see Supplementary file S4 for the statistical test). This led us to explore the mechanisms that could explain the over-representation of terminal indels. Two documented mechanisms have been proposed to drive terminal indels: gene fusion/fission (Riley and Labedan 1997; Yanai et al., 2001) and ‘intra-domain recombi-nation’ as exemplified by O’Sullivan et al. (2000).

A careful analysis of the syntenic contexts of the proteins reshaped by terminal indels reveals that 42% (38 out of 90) of these correspond to what we called a ‘straightforward fusion/fission’ (Table 3) and thus have been obviously rearranged by gene fusion/fission [see Supplementary material Table S2 for the KEGG (Kanehisa et al., 2004) and COG (Tatusov et al., 2000) annotations of the straightforward fusions/fissions]. An example of ‘straightforward fusion/fission’ is given in Figure 2 where gene HI0147 from H.influenzae corresponds to the straightforward fusion of genes VC1777 and VC1778 from Vibrio cholerae. The notion of ‘straightforward fusion/fission’ correlates well with a study of Yanai and co-workers (Yanai et al., 2002) suggesting that evolution by gene fusion involves an intermediate stage during which the future fusion components co-exist as juxtaposed but still distinct genes.

On the whole, 38 events out of 141 clearly correspond to gene fusions/fissions. Thus, it can be estimated that the contribution of gene fusion/fission to the evolution of multi-domain proteins is 27% (38/141). This is a lower bound. Indeed, we assumed here that none of all the ‘other terminal indels’ (52 = 90 − 38) is because of a gene fusion/fission event. Yet, a terminal indel which is not substantiated by a straightforward fusion/fission may be explained by a process involving gene fusion/fission. This point is addressed in the Discussion section.

Terminal repetitions are not explained by gene fusion/fission According to Andrade et al. (2001) ‘repeats are thought to arrive via intragenic duplication and recombination event’. Our results correlate well with this suggestion. Indeed, among the 34 cases of terminal domain repetitions, only 3 are because of straightforward fusions/fissions whereas 31 are not (data not shown). This indicates that domain repetitions do not mainly occur through gene fusions/fissions. It also demonstrates that our methodology (i.e. the way we collected our data set of homologous multi-domain proteins) is sound and correct.

DISCUSSION

The contribution of gene fusion/fission events to the evolution of bacterial multi-domain proteins is upper-bounded by 64% As outlined before, we showed that 42% of terminal indels are detected as straightforward fusion/fissions. The importance of this percentage led us to design a scenario by which the terminal indels that do not correspond to straightforward fusions could nevertheless be explained by a process of gene fusion. The scenario
is based on the three-step procedure depicted in Figure 3. As shown in Table 1, ~40% of the domain teams host an ‘inserted gene’, i.e. a gene coding for a protein, the domains of which do not belong to the syntenic stretch. This is in agreement with the observation that the structure of bacterial genomes is highly dynamic (Casjens, 1998; Tillier and Collins, 2000; Omelchenko et al., 2003; Rocha, 2004). Therefore, a gene can easily be inserted into a syntenic genome stretch. If such a gene fuses with one of its neighbours, then no mark will remain to indicate that this terminal indel is the result of a gene fusion. In a similar way, a gene may be split into two parts and one part may be excised from the syntenic stretch; as in the case of fusion, no mark will remain to indicate that this terminal indel is the result of a gene fission. This suggests that some (or many) of the ‘other terminal indels’ may well be attributed to plain gene fusions/fissions, increasing the prevalence of this evolutionary process. As a consequence, talking about ‘domain shuffling’ might be misleading in many cases. Indeed, a majority of new domain architectures might be better explained by ‘gene shuffling’ followed by fusion events. That is to say, domains do not shuffle but genes do and after their shuffling, genes may eventually fuse.

Based on the scenario described above, an upper bound of the contribution of gene fusion/fission to the evolution of multi-domain proteins can be estimated by assuming that all the ‘other terminal indels’ are because of plain gene fusion/fission. This gives an upper estimate of 64% (90/141).

Our aim here is not to rule out other mechanisms as being contributors to evolution of bacterial multi-domain proteins. However, we believe that gene fusion/fission might be the major contributor. Riley and Labedan (1997) already suggested that any multi-domain proteins might be the result of gene fusion. Kummerfeld and Teichmann (2005) showed that fusion/fission are frequent events (fusion being four times more frequent than fission). However, to draw their conclusion, these two works rely on bases which are not as firm as it seems. Thus, for instance, Kummerfeld and Teichmann (2005) looked for domain architectures that are present as a single protein in at least one genome (composite form) and as a set of shorter proteins in other genomes (split form), irrespective of the
location on the genome of these shorter proteins. For these authors, these composite and split domain architectures represent orthologous proteins. In our opinion, this criterion is too loose while in our approach, the use of the syntenic context allows to establish an unambiguous connection between composite and split forms. Finally, note that a very recently published work (Weiner and Bornberg-Bauer, 2006) substantiates our analysis since it provides evidences showing that a particular class of multi-domain protein rearrangement, called circular permutation, probably evolved through gene fusion/fission.

Checking for sequencing errors in straightforward fusions/fissions

To fully assess the methodology used in the present study, we wondered whether fused/unfused genes could be the result of gene-prediction or sequencing errors (which would make our results irrelevant). In the case of bacterial genomes, the object of the present study, a false straightforward fusion could only be attributed to a sequencing error such as a nucleotide omission (Koonin and Galperin, 2003) leading to an artefactual frameshift. Thus, for each identified straightforward fusion in a domain team, we searched for the presence of a similar fused (resp. unfused) form in a set of closely related genomes (w.r.t. the taxonomy). Indeed, if each of the fused and unfused forms can be identified in several closely related genomes, the fusion is most unlikely to be the result of a sequencing error (Kummerfeld and Teichmann, 2005). It turned out that, according to the previous criterion, 71% (27 cases) of the straightforward fusions/fissions detected in this study are not spurious (see Supplementary Material Table S3 for the results of this analysis). Note that this analysis required the use of additional genomic sequences not listed in the bacterial sets.

ACKNOWLEDGEMENTS

The authors are grateful to Jean-Luc Ferat, Meriem El Karoui and to the members of ABI (University of Paris VI) for helpful discussions. The authors thank the two anonymous referees for their useful and relevant comments.

Conflict of Interest: none declared.

REFERENCES