BIOINFORMATICS ORIGINAL PAPER
Vol. 21 no. 8 2005, pages 1464–1471
doi:10.1093/bioinformatics/bti204

Structural bioinformatics

Convergent evolution of domain architectures (is rare)

Julian Gough
RIKEN Genomic Sciences Centre, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama 230-0045, Japan

Received on August 31, 2004; revised on November 24, 2004; accepted on December 2, 2004
Advance Access publication December 7, 2004

ABSTRACT

Motivation: In this paper, we shall examine the evolution of domain architectures across 62 genomes of known phylogeny including all kingdoms of life. We look in particular at the possibility of convergent evolution, with a view to determining the extent to which the architectures observed in the genomes are due to functional necessity or evolutionary descent. We used domains of known structure, because from this and other information we know their evolutionary relationships. We use a range of methods including phylogenetic grouping, sequence similarity/alignment, mutation rates and comparative genomics to approach this difficult problem from several angles.

Results: Although we do not claim an exhaustive analysis, we conclude that between 0.4 and 4% of sequences are involved in convergent evolution of domain architectures, and expect the actual number to be close to the lower bound. We also made two incidental observations, albeit on a small sample: the events leading to convergent evolution appear to be random with no functional or structural preferences, and changes in the number of tandem repeat domains occur more readily than changes which alter the domain composition.

Conclusion: The principal conclusion is that the observed domain architectures of the sequences in the genomes are driven by evolutionary descent rather than functional necessity.

Contact: gough@supfam.org

INTRODUCTION

A domain is the smallest unit of evolution by the definition from the SCOP (Murzin *et al*., 1995) database of known protein structures. Small proteins consist of a single domain, and some larger proteins consist of more than one domain. A part of a protein is only considered a domain in its own right if it is observed elsewhere in nature on its own or in combination with different partner domains. Domains with structural, functional and sequence evidence for a common evolutionary ancestor are classified within the same superfamily in SCOP. The domain architecture of a protein is described by the order of the domains and the superfamilies to which they belong. The repertoire of architectures present in the genomes has arisen by the duplication and recombination (Miyata and Suga, 2001; Wagner, 2001; Ohno, 1970) of the ancestral superfamily domains (Chothia *et al*., 2003; Qian *et al*., 2001), often forming larger multi-domain proteins (Rossmann *et al*., 1974).

The SUPERFAMILY (Gough *et al*., 2001; Madera *et al*., 2004) database of hidden Markov models (HMMs) (Krogh *et al*., 1994; Eddy, 1996; Hughey and Krogh, 1996) representing all proteins of known structure assigns SCOP superfamily domains to all genome sequences, achieving a coverage of about half of all amino acids in each of the 227 (at the time of writing) complete genomes. From these domain assignments, the domain architecture of each sequence is derived as described previously (Vogel *et al*., 2004). All data are available from http://supfam.org. The evolutionary-based domain definitions, for which three-dimensional (3D) structures are required, are necessary to draw the conclusions from this work. Including protein domains from other databases, such as Pfam (Bateman *et al*., 2000), SMART (Ponting *et al*., 1999) or InterPro (Apweiler *et al*., 2001), for which there are no 3D structures, and hence no confirmed evolutionary relationships, would make some improvement to the coverage but at the cost of complicating the arguments. Other databases do not conform to the SCOP definition (above) of evolutionary relationships, which is the basis for the conclusions.

The primary question which is addressed here, is to what extent the architectures observed in the genomes are due to functional necessity or due to evolutionary descent, i.e. to what extent genes' stringent selective requirements have led to identical architectures on multiple occasions. Convergent evolution is defined here as more than one independent evolutionary event (recombination) leading to the same domain architecture in different genomes. If the shuffling of domains is functionally driven then we expect to find a great deal of evidence of convergent evolution, since the same architecture would be arrived at independently in several different genomes. A failure to detect convergent evolution points to evolutionary descent being the explanation for the observed presence of architectures in the genomes. This point of view is supported by previous analyses of Rossmann domains (Bashton and Chothia, 2002) and domain-pairs in genomes (Apic *et al*., 2001). The evolution of the domains with respect to each other (Copley and Bork, 2000; Rost, 2002), or with respect to networks (Amoutzias *et al*., 2004; Conant and Wagner, 2003a) is not considered here.

It is important to address this question since any future work on domain architectures, and some current research (Vogel *et al*., 2004), depends upon or is relevant to it (Amoutzias *et al*., 2004; Ranea *et al*., 2004). It is also important in its own right as it provides an insight to the driving forces behind the evolution of duplication and recombination; it is an essential piece of the complete puzzle of protein and genome evolution (Snel *et al*., 2002; Kunin and Ouzounis, 2003). Without understanding this phenomenon it is not possible to draw watertight conclusions based upon the observation of architectures present in the different genomes.

This paper first identifies possible cases of convergent evolution using phylogeny, sequence length, similarity and mutation rates of domains. It then examines the list of candidates for cases which have been falsely identified due to errors in the assignment of domain architectures from the SUPERFAMILY database. Finally, plausible evolutionary scenarios of convergent evolution are sought for the remaining candidates. These analyses together allow us to assess the extent of convergent evolution.

SYSTEMS AND METHODS

In the work presented here, we used the 78 441 sequences which had 3899 completely assigned domain architectures from 62 genomes in version 1.63 of the SUPERFAMILY database, i.e. those with incomplete or partial sequence coverage were not included.

The occurrence of architectures in the genomes is considered in a simplistic way. If an architecture is present in one genome and not in another, it is considered as having been lost by one or gained by the other. This is not to be confused with gene loss (Krylov *et al*., 2003) or acquisition, since the modification of a gene leading to a change in architecture, may be seen as the apparent loss of one architecture and gain of another. These are the terms in which architectures are considered in this paper.

Phylogenetic tree

The genomes used were chosen because their phylogeny has been previously examined and is reasonably well understood, as shown in Figure 1. The bacterial tree was grouped together and taken from Wolf *et al*. (2002), and the coelomate organization was also taken from Wolf *et al*. (2004) although this remains controversial (Copley *et al*., 2004). Plants and fungi are considered out-groups, with the root placed between single and multicellular organisms; the eukaryote sub-tree is not rooted.

Candidates for convergent evolution

Phylogenetic grouping The proteins with each domain architecture were first classified into phylogenetic groups in the following way, which is similar to that in previous work on horizontal transfer (Galperin and Koonin, 2000; Gaasterlan and Ragan, 1998). If an architecture is observed in the overwhelming majority of genomes belonging to a particular branch (or clade) of the phylogenetic tree, then it is almost certain that the architecture was inherited from the highest shared node of that branch. For a given architecture, the genomes were classified into phylogenetic groups sharing a common node as near the root of the tree as possible, satisfying the criterion that at least five out of every six contained the architecture.

Any architecture that forms more than one distinct phylogenetic group is a candidate for convergent evolution. The two explanations alternative to convergent evolution are horizontal gene transfer (Koonin *et al*., 2001; Kurland *et al*., 2003) and gene loss. Figure 2 shows that both of these cause the true evolutionary group descending from a common ancestral architecture to be split into more than one observed phylogenetic group. To ascertain whether, for any given architecture, phylogenetic groups should be joined to form a single evolutionary group, two different principles were used (described below). What they both make use of is the fact that sequences of architectures with phyletic patterns arising from gene loss and horizontal transfer share a common ancestor, having been created by one evolutionary event. Therefore, they will share characteristics, while sequences which independently evolved the same architecture will not.

Sequence length and similarity If two proteins come from the same complete-architecture evolutionary ancestor, then they will have a closer homology than the components of non-related proteins sharing the same architecture. Domains belonging to the same superfamily in SCOP may have diverged in sequence beyond the point where they share significant sequence similarity. Independent evolutionary events leading to the same architecture have combined the same superfamily domains in the same order, but may not have combined members from each superfamily that share high-sequence similarity.

Although the same order of domains is observed, any two random evolutionary events leading to the same architecture may not have chosen domains of the same length, or more importantly have stitched them together in the same places, with the same linking sequence or amount of truncation.

Architectures which form distinct phylogenetic groups (see above) were compared to see if they shared sequences with high-sequence similarity across their entire length, indicating that convergent evolution is unlikely. This was done conservatively to avoid the possibility of falsely eliminating candidates for convergent evolution. BLAST (Altschul *et al*., 1990) was used with local scoring and an E-value threshold of $E < 10^{-5}$. Local scoring chooses the best local alignment and will not cover the whole sequence unless there is a good match across the whole. An alignment was only accepted as covering the whole sequence if it included $(n+1)/(n+2)$ of all residues, where 'n' is the number of domains in the architecture.

Domain mutation rates The individual domains belonging to two proteins sharing the same complete-architecture ancestor will have diverged for the same length of time and in the same environment as each other (Conant and Wagner, 2003b). In two proteins which have convergently evolved the same architecture, the component domains will have a different evolutionary history from each other. Groups were linked to each other when sharing a similar protein, where the sequence identity of the corresponding domains in the similar pair of proteins varies by not more than 5% across all domains in the architecture. The sequence identity of the domain-pair (from the corresponding position in a pair of sequences) was calculated from alignments generated using SUPERFAMILY HMMs. Any sequence with domain pairs sharing <30% sequence identity were not linked, since sequence identity becomes inaccurate at low percentages, although using the HMM alignment greatly increases accuracy.

Architectures of the same composition Given several copies of the same architecture in several genomes, it can be difficult to demonstrate that they do not come from a common ancestor, and therefore the result of convergent evolution. To juxtapose this point of view let us consider architectures that contain the same domains but not in the same order. Since we know that they do not come from a common ancestor this does not help us to answer the question of vertical descent, but it allows us to examine another form of convergent evolution. Most importantly, we can compare the observed rate of this type of convergent evolution to that which we predict for similar architectures to see if they are the same.

All of the sequences were grouped based upon domain content and number, but regardless of organization. Where a group contains more than one architecture (ordering of the same domains) then we know that more than one evolutionary event has led to the same domain content. No phylogenetic trees or other grouping is required since two different architectures do not share a common ancestor. The number of cases of this type of convergent evolution, which would be expected by random domain shuffling can easily be calculated by randomly generating dummy architectures with the same number of domains and frequency as the observed architectures, and repeating the analysis.

Analysis of candidates

The aim of the candidate selection (above) was not to produce a high-quality list of sequences that have evolved by convergent evolution, but rather to eliminate those sequences with architectures that have most probably not evolved by convergent evolution. Therefore, the resulting candidate set is expected to contain many members that are not examples of convergent evolution. Here, we analyze the cases of convergent evolution leading to the same architecture with the same ordering of domains.

Errors in architecture assignment The SUPERFAMILY database is designed to operate at an error rate of <1%, however owing to the statistical nature of sequence comparison methods, there will be some false assignments leading to incorrect domain architecture identification. Furthermore, there will be more cases where the sequences of some domains have diverged beyond the point of recognition. A failure to detect a domain in one case which is successfully detected in another will again lead to an incorrect domain architecture.

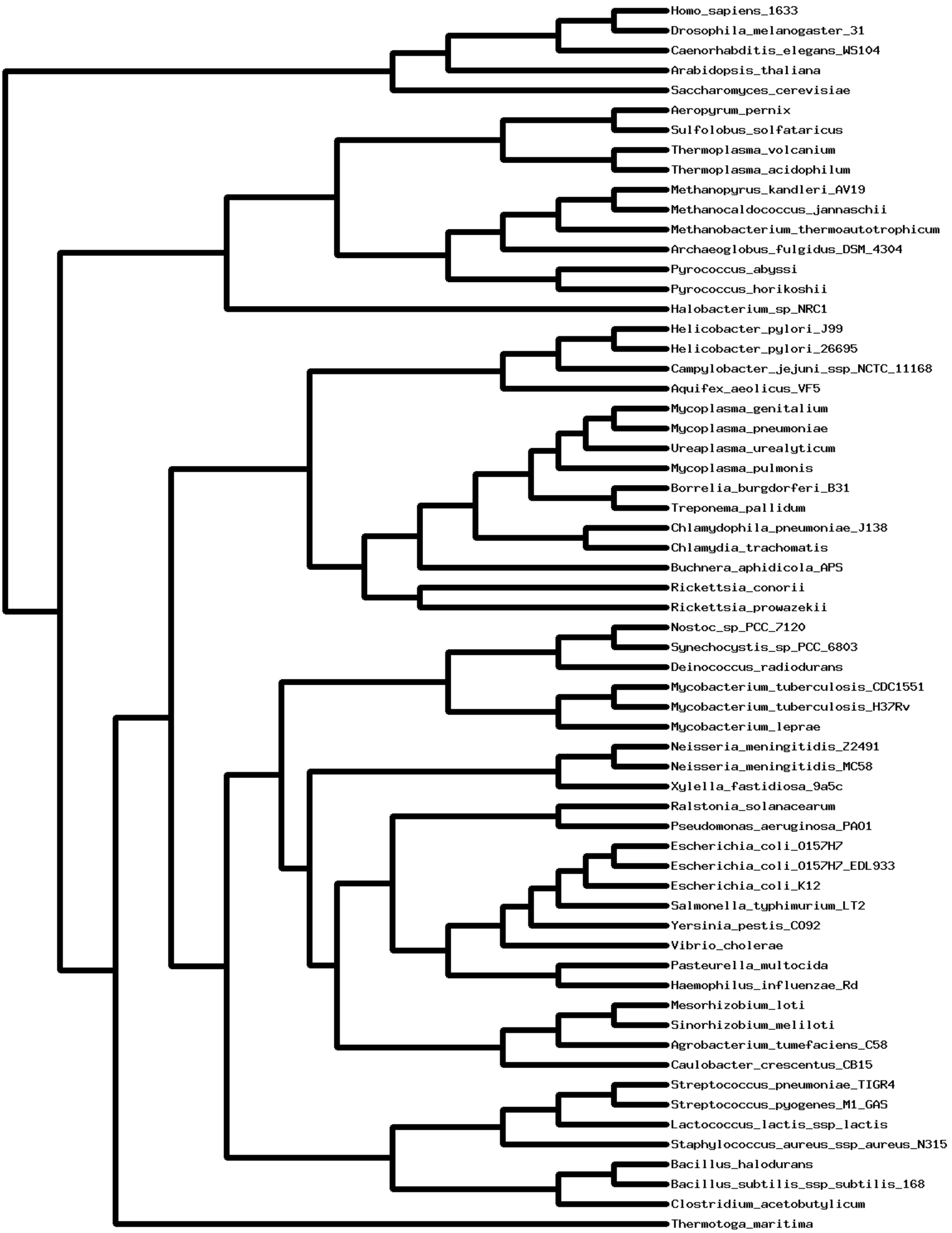

Fig. 1. Schematic representation of the unrooted phylogenetic tree of 62 genomes (the distances are arbitrary). The tree was drawn using scripts from http://supfam.org/treedraw/tree.html

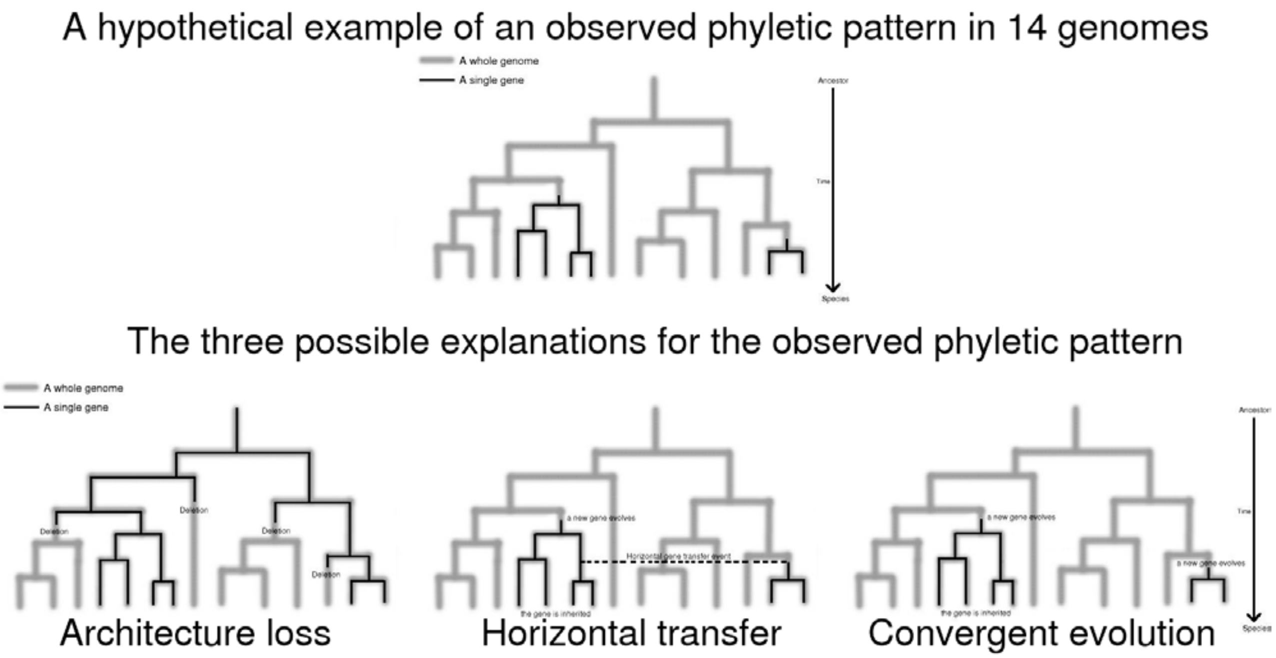

Fig. 2. A hypothetical example of an observed phyletic pattern of genomes for a given domain architecture, plus the three possible evolutionary explanations.

A single false domain assignment could lead to a one-off architecture observation in a distantly related genome not in the legitimate phylogenetic group. Hence, all candidate architectures were examined for cases where the candidate would be eliminated if it were not for a singleton sequence outside the main group. If one of the domains in the singleton sequence has a borderline score with an E-value, $E > 10^{-4}$, yet all of the domains in all of the sequences in the main group have good scores with an E-value, $E < 10^{-8}$, then it was decided that the assignment is most likely to be false. Therefore, convergent evolution is not the explanation.

Negative errors, i.e. undetected domains could lead to a sequence having incomplete coverage with the offending domain missing. As stated above, these incompletely covered sequences were excluded from the main analysis; we reconsidered them for the candidate architectures. That is, we searched the previously excluded set of sequences, which are not completely covered from N to C termini by domain assignments, to find sequences that had the same architecture as one of our candidates, except that they have an empty unassigned space of the correct length where the right domain would make up a complete architecture which would be the same as the candidate. We looked for sequences with the right architecture, but for one missing domain, in the genomes which would 'fill in the gaps' in the tree allowing a single phylogenetic grouping. If these sequences in fact have the same architecture, but were not identified, then their phyletic patterns are not caused by convergent evolution. To eliminate cases of missing more than one domain from the complete architecture of the sequence, or missing another domain of different length, the length of the unassigned region was examined. Unless the length of the unassigned region was in the interval $L_{\max} + 5 + (L_{\max} - L_{\min})/4 > L > L_{\min} - 5 - (L_{\max} - L_{\min})/4$, with $L_{\max}$ and $L_{\min}$ being the greatest and least observed lengths of the domain in fully assigned sequences, then they were rejected. If the unassigned region of a sequence is far beyond the extremes of observed length of the expected domain, then it is unlikely that it is present and undetected in that region.

Evolutionary scenarios It is very difficult to find substantial evidence of a case of convergent evolution of domain architectures. However, where circumstantial evidence can be found such as a plausible scenario for a mechanism, for example gene fusion or fission, it would appear more likely to be a real case.

For a given architecture, by examining the genomes that would be required to make up a single complete phylogenetic group but are missing the said architecture (as with 'negative errors' above), we were able to see if the phyletic pattern may be explained by independent events of the same nature:

- If the genomes missing the architecture have two sub-components of the architecture, at least one of which is not itself present in those genomes that do have the architecture, then independent gene fusion events could explain the observed phyletic pattern.
- If the genomes missing the architecture have another larger protein of which the architecture is a sub-component and the other sub-component(s) of the larger protein are present in the genomes that do contain the architecture, then independent gene fission events could explain the observed phyletic pattern.
- If the genomes missing the architecture contain architectures that are similar, but with a different number of tandem repeats of one of the domains in the architecture, then independent tandem domain gain/loss events could explain the observed phyletic pattern.

IMPLEMENTATION

Candidates for convergent evolution

The mono-domain architectures, by SCOP definition at the superfamily level (see Introduction section) share a common evolutionary ancestor, and as such cannot have arisen by convergent evolution. The question of whether or not in a few outlying cases domains belonging to the same superfamily may have arisen by convergent evolution of domains is not addressed because they would not have arisen by domain shuffling. It is not relevant to the convergent evolution of architectures. Thus, the mono-domain architectures can be used as a control.

Table 1. Grouping of the 3899 domain architectures by phylogenetic clade

Groups	Mono-domain architectures	Multi-domain architectures	All architectures
1	280	2240	3098
2	131	315	315
3	83	128	128
4	53	79	79
5	38	55	55
>5	273	224	224

Each architecture is present in one or more genomes; these genomes are grouped together based upon membership of the same phylogenetic clade; each architecture therefore contains one or more groups. Each column of this table shows how many architectures contain the number of groups corresponding to the first column. For example, there are 315 multi-domain architectures whose genomes group into two phylogenetic clades. Mono-domain architectures share a common ancestor, so the mono-domain column is shown as a control. In the final column all mono-domain architectures are listed as having one group, since regardless of the observed phylogenetic distribution they are known from structure to have descended from a single common ancestor via one process or another.

Table 2. Further grouping (of Table 1) by sequence length and similarity

Groups	Mono-domain architectures	Multi-domain architectures	All architectures
1	431	2652	3510
2	201	268	268
3	77	60	60
4	58	28	28
5	38	15	15
>5	53	18	18

This table is of the same format as Table 1. In this case however the data shown are the results of applying further grouping based upon sequence length and similarity in addition to the simple phylogenetic grouping shown in Table 1. For some architectures this will group together sequences that share a common evolutionary ancestor despite belonging to genomes of apparently different clades. They may belong to genuinely different clades due to horizontal gene transfer, or appear to belong to different clades because disappearance of the architecture from some genomes affects the distribution and obscures the phylogenetic history.

Phylogenetic grouping The results of the phylogenetic grouping are shown in Table 1 where 3098 out of 3899 architectures form a single phylogenetic group, which means that they are presumed not to be involved in convergent evolution. The mono-domain architectures are assumed to form all single groups (as reasoned above) but the first data column is included as a control; this shows that the phylogenetic grouping is conservative, and will not group together many things which should be. However, the phylogenetic grouping is more effective on multi-domain architectures, because on average they are more recently evolved and have undergone less deletion and horizontal transfer; groups of mono-domain proteins from the set contain sequences, which have on average diverged to ∼10% identity, whereas multi-domain proteins have diverged on average to only ∼30% sequence identity.

Sequence length and similarity The analysis described in the Systems and Methods section was applied to the groups obtained from the phylogenetic grouping (above), and is shown in Table 2. Once again the mono-domain architectures are shown as a control, yet in

Table 3. Grouping by domain mutation rates compared to, then combined with, results from Table 2

Groups	Domain mutation rates	Sequence length and similarity	All grouping applied together
1	3423	3510	3527
2	276	268	255
3	94	60	58
4	41	28	27
5	18	15	17
>5	47	18	15

The format of this table is similar to Tables 1 and 2. In this case the results from grouping by domain mutation rate is shown in the first data column, and the results from grouping by sequence length and similarity is shown in the second data column (the same as the final column in Table 2). The third data column in this table shows the results of applying together both of the methods for grouping from the previous two columns. Note that the numbers in all three columns are similar, indicating that the two methods independently concur with each other.

the final column, all mono-domain architectures are put in a single phylogenetic group. This analysis is still more conservative than phylogenetic grouping with respect to the mono-domain control. This is not surprising since many domains diverge beyond the point of recognition by simple pairwise sequence comparison methods such as BLAST.

Domain mutation rates This analysis described in the Systems and Methods section was applied to the groups obtained from the phylogenetic grouping, and to the groups obtained in the previous section separately. Thus, Table 3 is different to the previous two, and no control is shown because domain mutation rates cannot be compared unless there is more than one domain. Although the results obtained here (column 2) are similar to those obtained in the previous table (column 3), combining them (column 4) makes very little difference. The grouping is again conservative since as multi-domain proteins diverge, their domains having different functions and structure, undergo different mutation rates due to different selective pressures.

Parameters used for analysis Although the effect of varying the chosen parameters for the analysis was investigated in depth, it is not presented here in full detail. However, the numbers of candidates as a fraction of the total are little affected by minor changes in the chosen parameters, which are designed to be conservative. Changing the cut-off for 5 out of 6 genomes for phylogenetic grouping changes the number of architectures forming a single group by a maximum of only 0.004% (10–13 architectures) when considering a cut-off of either 9 out of 10 genomes in a clade (stricter), or 4 out of 5 (less strict). Changing the required coverage of Blast local matches from $(n+1)/(n+2)$ to $n/(n+1)$ or $(n+0.5)/(n+1.5)$ changes the number of architectures within a single group by 1.2 and 0.6% respectively. Changing the percentage difference in domain mutation rates from 5 to 3% causes the total number of architectures forming a single group by 0.97%, and changing the lowest sequence identity considered accurate from 30 to 25 or 35% changed the total number by 1.0 and 0.9%, respectively.

Architectures of the same composition In the 78 441 sequences in this study, we found 113 cases where different domain architectures had evolved with the same composition of domains. These cases do not share a common ancestor, yet have evolved the same number of domains belonging to the same superfamilies. As described in the Systems and Methods section, we calculated that an average of 138.6 cases and a SD of 10.9 (in 50 trials) would arise by random domain selection. The number from this naive calculation is in fact higher than the observed quantity, yet close to it, suggesting that the domain shuffling leading to the observed architectures is indeed random.

Furthermore, both the observed number of cases and that predicted by random domain shuffling are of a similar magnitude to the predicted number of cases for convergent evolution leading to similar architectures, which supports the final estimates stated in the discussion.

Analysis of candidates

The remaining 372 candidate architectures from the final column of Table 3 account for $\sim$10% of the architectures, and 7% of the total number of sequences. These candidates alone are considered below.

Errors in architecture assignment Of the 372 candidate architectures 217 have phyletic patterns with a singleton genome, such that there would only be one phylogenetic group if that genome were falsely accredited with the architecture in question. However, only 30 of these had domains with borderline scores fitting the criteria described in the Systems and Methods section, and are probably false. This is roughly what would be expected at an error rate of $<$1%.

Furthermore, 81 architectures from the candidate set were identified as having phyletic patterns, which would form a single group, if domains were failed to be detected in regions of the correct expected size (as described in the Systems and Methods section), in sequences from the missing genomes needed to make up the clade. This is a larger number since domain assignments from SUPERFAMILY have more false negatives than false positives.

Evolutionary scenarios The candidate architectures were analysed for possible evolutionary mechanisms to explain their occurrence. We found 49 cases where the architectures which would form a single phylogenetic group had different numbers of tandem repeats of the same domain, causing separate groups to be identified by the analysis. However, only seven of these included only tandem repeats of at least two domains. These are potential examples of convergent tandem duplication. We found in addition to these that there were 10 examples of potential convergent gene fission and only a single example of potential convergent gene fusion of non-tandem domains.

So all in all we found 59 candidates that have a plausible evolutionary explanation of either convergent tandem duplication, gene fusion or gene fission. These were examined in detail.

In the case of the tandem repeat architectures, we found that in more cases than average, sequences with the given architecture were found in genomes in many different phylogenetic groups. Of the 49 architectures, 22 were in three or more phylogenetic groups and 6 were in more than five groups. When compared to the last column of Table 3 we see that this is a disproportionate amount. Phyletic patterns with genomes forming many groups are more consistent with a high rate of gene loss, than with the alternative low likelihood of multiple repeated convergent evolution events required to explain the observation. This presupposes the conclusion that convergent evolution is not extremely common. This, in combination with the large number of candidates in comparison with those for fusion and fission, suggests that the number of tandem repeats in an architecture evolves more rapidly, and is less functionally constrained, than changes involving loss or gain of different (non-repeat) domains. The 27 architectures (223 sequences) that could be explained by two independent, yet similar, evolutionary events were found to utilize a wide variety of domains in architectures of varying length, function and distribution across the kingdoms of life. There is no discernible preference or feature of the set.

Likewise the 10 fusion/fission candidates were varied, yet involved human or *Arabidopsis* genomes in all but one case. Two of the ten cases involving human [ENSEMBL (Hubbard *et al*., 2002) version 16.33] disappear in the latest release of the genome (19.34a). Another architecture has an unassigned region at the N-terminus which could be a missing domain or a tail or extension to the first assigned domain; the apparent case of convergent evolution is due to inconsistency of the architecture assignment algorithm described by Vogel *et al*. (2004). The remaining seven cases of fusion/fission (119 sequences) appear to be genuine, notwithstanding alternative splice variants, sequencing and gene prediction errors:

- An architecture which consists of a P-loop domain, and the N- and C-terminal subunits of F1 ATP synthase occurs in eukaryotes, yet *Arabidopsis* (and rice) have lost the C-terminal subunit. Since a group of bacteria share the same architecture as *Arabidopsis*, it appears to be an example of convergent evolution.
- A common architecture variant in eukaryotes is a string of tandem immunoglobulin domains, with some fibronectin domains. The numbers of tandem repeats of both domains varies both within and between genomes, but in human and *Caenorhabditis elegans* there exists an architecture with no fibronectin domains. There may exist splice variants not predicted by the gene-prediction programs which would negate this case.
- Two remotely related archaebacteria (*Sulfolobus solfataricus* and *Thermoplasma acidophilum*) have fused an additional adenine nucleotide alpha hydrolase-like domain with an existing domain paired with a Rossmann-like domain. The other genomes in the clade contain the two unfused architectures.
- A P-loop domain and an EF-hand appear in combination in two eukaryotes: *Arabidopsis* and *C.elegans*. They appear to have been created by independent evolutionary events, both via loss of a second EF-hand domain sandwiching the P-loop.
- Two evolutionarily distant bacteria (*Bacillus halodurans* and *Salmonella typhimurium*) have truncated two-domain variants of a three-domain architecture which is present in almost all bacteria: a Rossmann-like domain sandwiched between two thiamin diphosphate-binding domains.
- A glutathione synthetase ATP-binding domain in combination with a preATP-grasp domain is observed in human and *C.elegans* (disappears from *Arabidopsis* in latest release of genome) and most bacteria. It would appear that eukaryotes have expanded this two-domain architecture, all of them having variants with an additional SCOP ‘Rudiment hybrid motif’ domain at the C-terminus, sometimes followed by other domains. Human and *C.elegans* each have at least one sequence which has reverted to the pair-form.

- A P-loop domain appears in combination with a translation protein in *Arabidopsis* and *T.acidophilum*. All but two genomes have the same architecture but with a EF–Tu/eEF–1alpha/eIF2–gamma C-terminal domain. The two genomes seem to have convergently evolved architectures (in addition) which have lost the C-terminal domain.

DISCUSSION

It is clear that the vast majority of sequences in the genomes have domain architectures which have arisen by evolutionary descent rather than due to functional necessity (see the Introduction section). In short, convergent evolution of domain architectures is rare. This work does not claim to provide irrefutable proof with exhaustive coverage, but it does give a clear overview and present a small number of strong cases.

The ∼2% of sequences with architectures forming three or more distinct phylogenetic groups may be explained either by a higher rate of deletion for this architecture, or by multiple parallel convergent evolution events. Disregarding the unlikely latter scenario, 5% of the sequences remain as potential candidates for convergent evolution. Taking into account the false positives in the SUPERFAMILY assignments does not make much difference, but in combination with the potential false negatives the candidates lower to 4% of all sequences. In reality most candidates will not be true cases. Including the tandem repeat variants, 0.4% of the sequences were shown probably to be true cases, notwithstanding gene sequencing and prediction shortcomings. We conclude that the upper and lower bounds are 4 and 0.4% respectively, but that for reasons discussed below, we expect the actual proportion to be close to the lower bound.

The methods for candidate selection are conservative, and although it is possible in some cases to eliminate the possibility of convergent evolution, it is very hard to rule out loss of architectures as an explanation. Architectures which came about a long time ago in evolutionary history may have diverged beyond the point of recognition, and these will also have had more time to undergo losses in some genomes. It is most likely that these losses and divergence account for most of the candidate 4%. Fallibility of the architecture assignment, and ENSEMBL gene predictions, accounted for 3 out of 10 candidates that were closely scrutinized to be falsely detected; this could apply to some of the tandem repeat candidates as well. Many false candidates may not have been eliminated because of sequencing and gene prediction errors, in particular with regard to splice variants that may account for more expressed proteins than are currently predicted by the genome projects (Zavolan *et al*., 2003; Okazaki *et al*., 2002).

These estimates for the rate of independent convergent evolution events producing the same architecture, are supported by the observed rate of convergent evolution leading to different architectures with the same domain composition. Furthermore this rate is no greater than that which would be expected by random domain shuffling, which further strengthens the argument that the observed architectures have not arisen as a result of functional necessity.

There are no discernible patterns or characteristics of convergently evolved domain architectures, which is in keeping with a random model for mutations, duplication/ recombination and gene fusion/fission events. Simply put, the examples of convergent evolution appear to have occurred by chance and without preference for function or structure. The sample is however, too small to be conclusive.

Again, from a small sample, it appears that variations in numbers of tandem repeats of the same domain, evolve faster than other forms of recombination. Probably as a result of this, there are more examples of convergent evolution involving changing numbers of repeats than involving gain or loss of different domains. This makes sense from the points of view of both the mechanism and the function. Changing the number of repeat domains may require mutations over shorter genomic distances, and such changes may have a milder effect on the resulting function (and possibly structure) of the protein than changing completely different domains.

ACKNOWLEDGEMENTS

Thanks to Sarah Kummerfeld for discussing her work. This work is supported by RIKEN care of Dr Hayashizaki.

REFERENCES

Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.

Amoutzias,G.D., Robertson,D.L., Oliver,S.G. and Bornberg-Bauer,E. (2004) Convergent evolution of gene networks by single-gene duplications in higher eukaryotes. *EMBO Rep.*, **5**, 274–279.

Apic,G., Gough,J. and Teichmann,S.A. (2001) Domain combinations in archaeal, eubacterial and eukaryotic proteomes. *J. Mol. Biol.*, **310**, 311–325.

Apweiler,R., Attwood,T.K., Bairoch,A., Bateman,A., Birney,E., Biswas,M., Bucher,P., Cerutti,L., Corpet,F., Croning,M. *et al*. (2001) InterPro database, an integrated documentation resource for protein families, domains and functional sites. *Nucleic Acids Res.*, **29**, 37–40.

Bashton,M. and Chothia,C. (2002) The geometry of domain combination in proteins. *J. Mol. Biol.*, **315**, 927–939.

Bateman,A., Birney,E., Durbin,R., Eddy,S.R., Howe,K.L. and Sonnhammer,E.L. (2000) The Pfam protein families database. *Nucleic Acids Res.*, **28**, 263–266.

Chothia,C., Gough,J., Vogel,C. and Teichmann,S.A. (2003) Evolution of the protein repertoire. *Science*, **300**, 1701–1703.

Conant,G. and Wagner,A. (2003a) Convergent evolution of gene circuits. *Nat. Genet.*, **34**, 264–266.

Conant,G.C. and Wagner,A. (2003b) Asymmetric sequence divergence of duplicate genes. *Genome Res.*, **13**, 2052–2058.

Copley,R.R. and Bork,P. (2000) Homology among (betaalpha) (8) barrels: implications for the evolution of metabolic pathways. *J. Mol. Biol.*, **303**, 627–641.

Copley,R.R., Aloy,P., Russell,R.B. and Telford,M.J. (2004) Systematic searches for molecular synapomorphies in model metazoan genomes give some support for Ecdysozoa after accounting for the idiosyncrasies of *Caenorhabditis elegans*. *Evol. Dev.*, **6**, 164–169.

Eddy,S.R. (1996) Hidden Markov models. *Curr. Opin. Struct. Biol.*, **6**, 361–365.

Gaasterlan,T. and Ragan,M.A. (1998) Microbial genescapes: phyletic and functional patterns of ORF distribution among prokaryotes. *Microb. Comp. Genomics*, **3**, 199–217.

Galperin,M.Y. and Koonin,E.V. (2000) Who's your neighbour? New computational approaches for functional genomics. *Nat. Biotechnol.*, **18**, 609–613.

Gough,J., Karplus,K., Hughey,R. and Chothia,C. (2001) Assignment of homology to genome sequences using a library of hidden Markov models that represent all protein of known structure. *J. Mol. Biol.*, **313**, 903–919.

Hubbard,T., Barker,D., Birney,E.A., Cameron,G., Chen,Y., Clark,L., Cox,T., Cuff,J., Curwen,V., Down,T. *et al*. (2002) The Ensembl genome database project. *Nucleic Acids Res.*, **30**, 38–41.

Hughey,R. and Krogh,A. (1996) Hidden Markov models for sequence analysis: extension and analysis of the basic method. *Comput. Appl. Biosci.*, **12**, 95–107.

Koonin,E.V., Makarova,K.S. and Aravind,L. (2001) Horizontal gene transfer in prokaryotes: quantification and classification. *Annu. Rev. Microbiol.*, **55**, 709–742.

Krogh,A., Brown,M., Mian,I.S., Sjolander,K. and Haussler,D. (1994) Hidden Markov models in computational biology. Applications to protein modeling. *J. Mol. Biol.*, **235**, 1501–1531.

Krylov,D., Wolf,Y., Rogozin,I. and Koonin,E. (2003) Gene loss, protein sequence divergence, gene dispensability, expression level, and interactivity are correlated in eukaryotic evolution. *Genome Res.*, **13**, 2229–2235.

Kunin,V. and Ouzounis,C.A. (2003) The balance of driving forces during genome evolution in prokaryotes. *Genome Res.*, **13**, 1589–1594.

Kurland,C.G., Canback,B. and Berg,O.G. (2003) Horizontal gene transfer: a critical view. *Proc. Natl Acad. Sci. USA*, **100**, 9658–9662.

Madera,M., Vogel,C., Kummerfeld,S., Chothia,C. and Gough,J. (2004) The SUPERFAMILY database in 2004: additions and improvements. *Nucleic Acids Res.*, **32**, 235–239.

Miyata,T. and Suga,H. (2001) Divergence pattern of animal gene families and relationship with the Cambrian explosion. *Bioessays*, **23**, 1018–1027.

Murzin,A.G., Brenner,S.E., Hubbard,T. and Chothia,C. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, **247**, 536–540.

Ohno,S. (1970) *Evolution by Gene Duplication*. Springer-Verlag, New York.

Okazaki,Y., Furuno,M., Kasukawa,T., Adachi,J., Bono,H., Kondo,S., Nikaido,I., Osato,N., Saito,R., Suzuki,H. *et al*. (2002) Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature*, **420**, 563–573.

Ponting,C.P., Schultz,J., Milpetz,F. and Bork,P. (1999) SMART: identification and annotation of domains from signalling and extracellular protein sequences. *Nucleic Acids Res.*, **27**, 229–232.

Qian,J., Luscombe,M. and Gerstein,M. (2001) Protein family and fold occurrence in genomes: power-law behaviour and evolutionary model. *J. Mol. Biol.*, **313**, 673–681.

Ranea,J., Bucjan,D., Thornton,J. and Orengo,C. (2004) Evolution of protein superfamilies and bacterial genome size. *J. Mol. Biol.*, **336**, 871–887.

Rossmann,M.G., Moras,D. and Olsen,K.W. (1974) Chemical and biological evolution of nucleotide-binding protein. *Nature*, **250**, 194–199.

Rost,B. (2002) Did evolution leap to create the protein universe? *Curr. Opin. Struct. Biol.*, **12**, 409–416.

Snel,B., Bork,P. and Huynen,M.A. (2002) Genomes in flux: the evolution of archaeal and proteobacterial gene content. *Genome Res.*, **12**, 17–25.

Vogel,C., Berzuini,C., Bashton,M., Gough,J. and Teichmann,S.A. (2004) Supra-domains—evolutionary units larger than single domains. *J. Mol. Biol.*, **336**, 809–823.

Wagner,A. (2001) Birth and death of duplicated genes in completely sequenced eukaryotes. *Trends Genet.*, **17**, 237–239.

Wolf,Y., Rogozin,I.B. and Koonin,E.V. (2004) Coelomata and not Ecdysozoa: evidence from genome-wide phylogenetic analysis. *Genome Res.*, **14**, 29–36.

Wolf,Y.I., Rogozin,I.B., Grishin,N.V. and Koonin,E.V. (2002) Genome trees and the tree of life. *Trends Genet.*, **18**, 472–479.

Zavolan,M., Kondo,S., Schonbach,C., Adachi,J., Hume,D.A., Hayashizaki,Y., Gaasterland,T., Group,R.G. and Members,G. (2003) Impact of alternative initiation, splicing, and termination on the diversity of the mRNA transcripts encoded by the mouse transcriptome. *Genome Res.*, **13**, 1290–1300.