Alternative to Homo-oligomerisation: The Creation of Local Symmetry in Proteins by Internal Amplification

ARTICLE *in* JOURNAL OF MOLECULAR BIOLOGY · SEPTEMBER 2009

Impact Factor: 4.33 · DOI: 10.1016/j.jmb.2009.09.031 · Source: PubMed

Anne-Laure Abraham
French National Institute for Agricultural R…
18 PUBLICATIONS **223** CITATIONS

Joel Pothier
Pierre and Marie Curie University - Paris 6
34 PUBLICATIONS **475** CITATIONS

doi:10.1016/j.jmb.2009.09.031

J. Mol. Biol. (2009) **394**, 522–534

JMB

Available online at www.sciencedirect.com

ScienceDirect

ELSEVIER

Alternative to Homo-oligomerisation: The Creation of Local Symmetry in Proteins by Internal Amplification

Anne-Laure Abraham[1,2]*, Joël Pothier[1] and Eduardo P.C. Rocha[1,2]

[1] *Atelier de BioInformatique, Université Pierre et Marie Curie–Paris 06, F-75005 Paris, France*

[2] *Microbial Evolutionary Genomics, Institut Pasteur, CNRS, URA2171, F-75015 Paris, France*

Received 23 February 2009; received in revised form 11 September 2009; accepted 15 September 2009
Available online 19 September 2009

Edited by M. Sternberg

The biologically active state of many proteins requires their prior homo-oligomerisation. Such complexes are typically symmetrical, a feature that has been proposed to increase their stability and facilitate the evolution of allosteric regulation. We wished to examine the possibility that similar structures and properties could arise from genetic amplifications leading to internal symmetrical repeats. For this, we identified internal structural repeats in a nonredundant Protein Data Bank subset. While testing if repeats in proteins tend to be symmetrical, we found that about half of the large internal repeats are symmetrical, most frequently around a rotation axis of 180°. These repeats were most likely created by genetic amplification processes because they show significant sequence similarity. Symmetrical repeats tend to have a fixed number of copies corresponding to their rotational symmetry order, that is, two for 180° rotation axis, whereas asymmetrical repeats are in longer proteins and show copy number variability. When possible, we confirmed that proteins with symmetrical repeats folding as an *n*-mer have homologues lacking the repeat with a higher oligomerisation number corresponding to the rotation symmetry order of the repeat. Phylogenetic analyses of these protein families suggest that typically, but not always, symmetrical repeats arise in one single event from proteins that are homo-oligomers. These results suggest that oligomerisation and amplification of internal sequences can interplay in evolutionary terms because they result in functional analogues when the latter exhibit rotational symmetry.

Keywords: evolution; oligomerisation; genetics; symmetry; repeats

Introduction

Most proteins are biologically active only in the form of oligomers containing some sort of symmetry.[1] Such symmetrical structures often result from the homomeric association of elements that are not themselves symmetrical.[2] While the reasons for this pervasive symmetry remain speculative, several hypotheses have been proposed. First, the symmetrical state could be the lowest-energy state and thus provides more stability.[3] Second, symmetry provides a simple way of building oligomers with a defined number of elements and therefore avoids aggregation.[1] It has also been proposed that the folding of symmetrical structures faces fewer kinetic barriers.[4] Furthermore, simulations on randomly docked complexes show that the lowest energy typically corresponds to symmetrical interactions.[5] *Escherichia coli* proteins show an average oligomerisation state of ~4 and only a minority of proteins is found in monomeric form. In general, the single most frequent complex state of a protein might be a dimer, most frequently a homodimer with a one-symmetry rotation axis (60–70% of all known complexes).[6] Within the remaining complexes, there are yet 20% of homomeric interactions. Homotetramers are less frequent than homodimers (15–20%), while homotrimers, homohexamers and homo-octamers are even rarer.[1,7] A minor fraction of proteins is found in the form of very long polymers or higher-order oligomers. Hence, associations among proteins leading to symmetrical structures are thought to play key roles in biological systems.

*Corresponding author. Atelier de Bioinformatique–Université Paris 6, Boite courrier 1202, 4 place Jussieu, 75252 Paris cedex 05, France. E-mail address: annela@abi.snv.jussieu.fr.

Abbreviation used: ssb, single-stranded DNA binding.

Transient or assortative interactions between proteins require the existence of independent molecules. Hence, a protein or a complex that participates in different complexes is expected to be coded by independent genes to allow for such modularity. But why are monomers not replaced by longer molecules for the vast number of proteins that establish long stable interactions within a single homomeric complex? Several evolutionary hypotheses have been put forward to answer this question: (1) Nature shows abundant examples where building by accumulation of construction bricks is adaptive.[8] (2) Assuming a constant error rate and a faulty protein elimination mechanism, it could be more efficient to construct multiple small subunits than larger ones.[9] However, it is unclear if the removal of mistranslated proteins is quick enough to prevent the establishment of a misfolded complex, knowing that such associations often lead to negative dominant phenotypes.[10] (3) The possibility of associating and dissociating subunits creates a potential for function enhancement and regulation. (4) It has been argued that oligomerised proteins are more evolutionarily constrained and thus subject to more stringent selection.[2] However, among functionally equivalent objects, evolution often favours elements that can easily evolve to adapt over elements that do not tolerate mutations.[11] This is because the purge of deleterious mutations involves the elimination of individuals carrying them from the population. This leads to higher genetic load, and it is therefore deleterious in most circumstances.

Close repeats occur spontaneously at high rates in both eukaryotes and prokaryotes and may result in duplication of structural domains.[12,13] Amplifications can also arise by exon shuffling in eukaryotes,[14] where proteins are indeed three times more likely to contain internal repeats.[12] Around 14% of proteins have been found to have long internal repeats.[12] These repeats have important evolutionary roles and are present even in small bacterial genomes.[15–18] Accordingly, recent works have shown relatively high frequencies of domain gain, loss and duplication in proteins.[19–21] Yet, there has been little work on a direct consequence of such events: that homo-oligomers could be replaced by symmetrical internal structures created by intragenic partial duplications. We conducted a study to test this idea. First, we identified internal repeats within a nonredundant data bank of protein structures. We then classed them as symmetrical or asymmetrical. It must be emphasised at this stage that what we call a symmetrical repeat is a set of structural elements in a given protein (i.e., copies of a repeat) that can be superimposed with a low resulting RMSD after a given symmetry operation. Many of these structural repeats cannot be strictly symmetrical, since in general they do not have strictly identical sequences. They should thus be called pseudo-symmetrical. Yet, for simplicity, we put together symmetrical and pseudo-symmetrical repeats under the same term. We classed repeat-containing proteins according to their structural features, to separate α-rich proteins, very repetitive proteins and the group of other proteins more likely to show features resembling that of homo-oligomers. We then used the latter set to search for homologous proteins with different multiplicities of the repeat, that is, proteins with just one copy of the repeated motif, proteins with two copies of the repeat and proteins with higher number of copies. Naturally, proteins with just one copy of the motif do not have a repeat. We found that proteins with one single copy of the repeat tended to have a doubled state of homo-oligomerisation relative to proteins with two copies of the repeat. We then analysed the evolution of the families of proteins with elements containing and lacking the symmetrical repeat in a phylogenetic framework.

Results

Proteins contain long symmetrical repeats

We searched for structural repeats longer than 50 residues among 8657 protein structures of the Astral data bank. We focused on long repeats because these have extremely low likelihoods of arising by random assembling of residues. Repeats were identified with Swelfe,[22] which uses dynamic programming to find optimal repeated substructures while weighting matches according to the frequency of α angles in the Protein Data Bank (PDB). This allows downplaying the role of very frequent α angles involved in archetypical secondary structural elements such as α-helices or β-sheets. We found 172 proteins containing long structural repeats. They correspond to ~2% of the data set (cf. Supplementary Table 1). We included in our analysis proteins of the Astral data bank with less than 50% sequence identity among themselves.[23] This avoids making multiple hits among closely related structures. We kept entire proteins, and not only domains, for our analyses. If some families of folds were overrepresented in our data set, this could inflate or play down the number of repeats. To check for this effect, we identified the presence of internal repeats in the "one structure per family" data set of Astral. We found ~3% of proteins containing long structural repeats. This is close to the ratio found with the "less than 50% identity" data set in Astral. Among the 172 proteins containing repeats, there are 103 different folds. This clearly shows that our results are not dominated by one or a few folds being overrepresented in the data set.

Since we used very stringent length and similarity criteria to identify internal repeats, we investigated if we would have found more than 172 proteins with more typical significance thresholds. The default values of Swelfe, score >250 and relative RMSD <0.5 (see Materials and Methods), are estimated to conservatively result in a p value of 10^{-3}.[22,24] Using these parameters, we found internal structural repeats in ~1900 structures, that is, in ~22% of the set. We wish to test a very specific hypothesis and

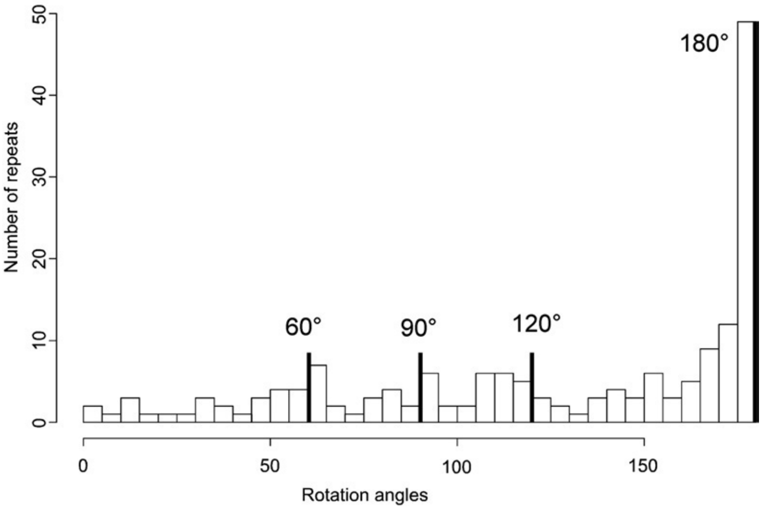

Fig. 1. Histogram of rotation angles allowing the superimposition of the two copies of the repeat for the 172 proteins containing repeats. Repeats with a 180° angle are very numerous and correspond to a 2-fold symmetry. There are also some small peaks at 60°, 90° and 120° that might correspond to 6, 4 and 3-fold rotation symmetry.

make a proof of principle. Hence, in the remaining analyses we preferred the use of the smaller but very reliable data set of long repeats, even if this represents only a small sample of the overall number of repeats.

We calculated the superimposition angle of the two copies of the internal repeats (see Materials and Methods and Supplementary Table 1). The histogram of rotation angles shows that rotations of ~180° vastly outnumber all others (Fig. 1). There are 61 repeats with a rotation angle between 170° and 180°, indicating that ~35% of long structural repeats have a 2-fold symmetry axis (C2). The 2-fold symmetry axis is pervasive among homodimers, which are the most abundant homo-oligomers. Hence, this large group of repeats is especially interesting to study in the framework of our hypothesis that sequence amplifications provide the opportunity to generate symmetrical structures analogous to that of homo-oligomers. We also found smaller peaks at rotation angles of ~120°, ~90°, and ~60°, which correspond to 3-, 4- and 6-fold symmetry (Fig. 1). While the number of proteins is low, the number of repeats with these rotational angles is higher than expected if distribution was uniform in the range 0–160° ($p=0.03$, $p=0.06$ for angles ≤170°, Wilcoxon tests). There are 89 proteins containing pairs of repeats (51% of the set) with verified rotation angles of 180°, 120°, 90° or 60°, showing that many of the long internal repeats are symmetrical under an axis of rotation. This is in agreement with our hypothesis that internal amplifications can give rise to symmetrical elements sharing structural resemblances with homo-oligomers, especially homodimers.

Classifying proteins according to their structure

We clustered proteins with internal repeats into three groups: α-helix-rich proteins, very repetitive proteins and other proteins (Fig. 2, Supplementary Table 1, see Materials and Methods for details). Since we suspected that these groups of proteins unveiled essentially different biological histories, we analysed them separately. By construction, the first group contains proteins with more than 85% of α angles in the range 40°–65°, which correspond to the angles found in α-helices. We compared the similarity score between the pairs of copies of the repeats in this group with those of the other two groups (Fig. 3): Scores corresponding to α-helix-rich repeats are significantly lower than the others

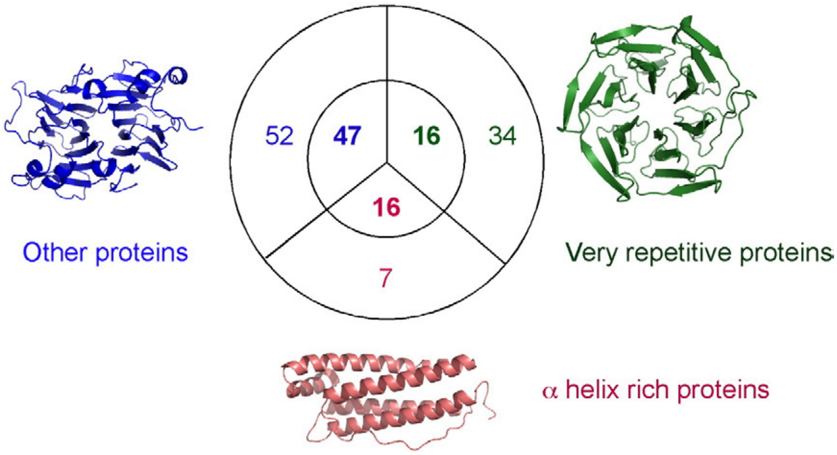

Fig. 2. Number of proteins with repeats that are symmetrical (2- to 6-fold) (inner circle in bold) or not (outer circle) in the three categories of repeats.

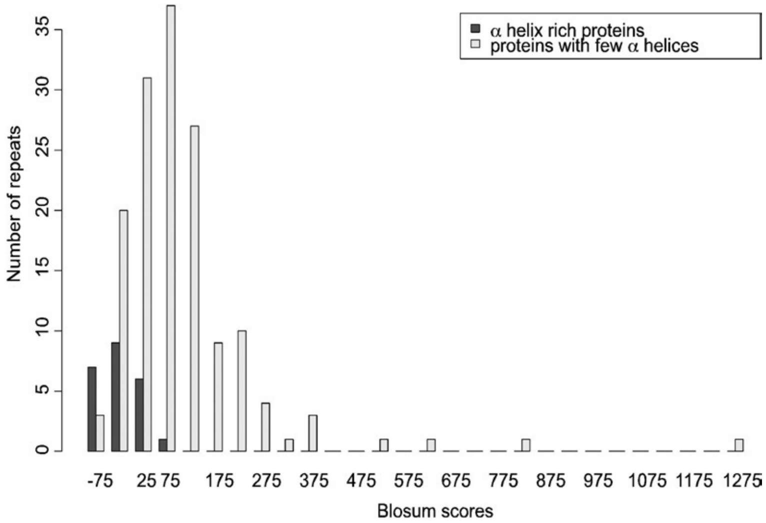

Fig. 3. Sequence similarity corresponding to structural repeats. The difference between the two groups is significant (Wilcoxon one-sided test, $p = 2.4\ 10^{-9}$).

($p = 2.4 \times 10^{-9}$, Wilcoxon one-sided test). Moreover, this group of proteins often showed negative sequence similarity scores between the two copies of the repeat. This means that the amino acid sequences of the two copies of the structural repeat are so dissimilar that they cannot be aligned meaningfully. This suggests that these structural elements are either very distant homologues whose sequences are saturated with changes or structural analogues resulting from convergent evolution.[25] α-Helices are very abundant because they result from a large variety of protein sequences and are thus particularly prone to convergent evolution. This class contains diverse functions. For example, the PDB entry 1cii corresponds to a protein of the colicin family. This ion-channel-forming protein kills bacterial cells by co-opting their active transport pathways and forming voltage-gate ion-conducting channels across the plasma membrane of the bacteria. The domain made up of two helices (160 amino acids long) that are nearly symmetrically repeated (rotation angle of 162°) enables the molecule to span the periplasmic space and contact simultaneously the outer and plasma membrane.[26] Other α-rich proteins include the botulinum neurotoxin type B (PDB entry 1s0e, with a symmetry angle of 175°),[27] which is a very potent toxin to humans and causes paralysis, and a bacterioferritin (1nf4, with a symmetry angle of 176°), which is able to store two iron ions.[28]

The second group of proteins with internal repeats contains very repetitive proteins. They were identified by visual inspection of structures containing at least six copies of the repeat. Repetitive proteins containing many α-helices are in the first group (α-helix-rich proteins). Some of these very repetitive proteins bind other proteins. For example, the Groucho protein (1gxr)[29] is a transcriptional co-repressor that interacts with DNA-bound transcription factors and histones. It contains a seven-bladed β-propeller WD40 repeat domain. The β-propeller is also found on a surface layer protein of *Methanosarcina* (1l0q).[30] Tropomodulin (1io0)[31] is a protein that blocks the elongation and depolymerisation of actin–tropomyosin. The domain that binds actin contains a leucine-rich repeat. Ankyrins (1n11)[32] link membrane proteins to the spectrin–actin cytoskeleton. The N-terminal domain binds the membrane and consists of 24 ANK repeats. Very repetitive proteins have probably not arisen by one single duplication event and, as expected, they rarely show 2-fold rotational symmetry (among the 16 proteins, only 3 have a 2-fold symmetry).

The remaining 99 proteins were put together in the third group, the majority of which contains structural repeats whose copies have significant sequence similarity. They have thus most likely arisen by internal amplifications of genetic material and we shall focus on their analysis. Among these 99 proteins we found 83 different folds, showing that this group contains very little redundancy in this respect. Around half (47) of these proteins contain repeats with 2-, 3- or 5-fold rotational symmetry (no 4- or 6-fold was observed in this group). These symmetries were visually checked in order to remove spurious symmetrical proteins: Some repeats are superimposed by a 60°, 90° or 120° rotation angle but do not correspond to 6-, 4- or 3-fold symmetry. We have tested if the repeats of this group correspond to duplications of Pfam domains, because these domains often correspond to functional units. Most repeats correspond approximately to the duplication of one Pfam domain; that is, each copy of the repeat overlaps along at least 70% of its length with one single domain (same or very similar name) (61 out of 99; Supplementary Table 2). In some cases, one single Pfam domain contains two copies of a repeat (20), and in others, a repeat copy contains two Pfam domains (2). The remaining 16 cases include proteins without domains or proteins with Pfam domains that do not overlap the repeats. There is thus a significant overlap between the structural repeats in this group and those of the Pfam domains, both among symmetrical and nonsymmetrical repeats. However, about a third of the repeats could not have been found by the analysis of Pfam domain duplications. Hence, our approach provides a different perspective on domain repeats from the ones previously published on Pfam domains. Contrary to the latter, it also allows the inclusion of information about the structural relative positioning of the different copies of a repeat within the protein.

No specific functional bias in proteins with symmetrical repeats

Is there any function overrepresented among proteins with repeats? To answer this question, we extracted the GO terms from Gene Ontology[33] of proteins with and without repeats. Using a previously published method, we then identified over- and underrepresented terms.[34] We only found one overrepresented term: Calcium-dependent phospholipid binding (Go term 5544) is overrepresented among proteins containing nonsymmetrical repeats. No term is over- or underrepresented in proteins with symmetrical repeats (at a $p<0.05$). This suggests that symmetrical repeats are not strongly associated with particular cellular components, biological processes or molecular functions. Naturally, the low number of proteins precludes the statistically meaningful identification of weak associations.

Many enzymes in the PDB are composed of several monomers: There are 7242 enzymes with known dimeric, tetrameric, hexameric or octameric quaternary structure, according to the PQS (Protein Quaternary Structure) data bank, among the 22,912 enzymes of the PDB (about 31%). Sometimes the association or dissociation of these monomers can regulate enzyme activity. In this case, a duplication overruling the necessity of oligomerisation and creating an enzyme in one single peptide chain could be selectively deleterious. We compared if enzymes were less frequent among proteins containing repeats (see Supplementary Table 3) or, within these, among proteins containing symmetrical repeats. EC numbers of enzymes were taken from the Enzyme Structures Database of PDBsum†. We assumed that structures with EC numbers were enzymes and that the remaining were not enzymes. Enzymes are underrepresented among proteins containing repeats ($p<0.005$, χ^2 test). However this difference is due to proteins in the group of highly repetitive proteins ($p=0.02$) and to some extent to the group of α-helix-rich proteins ($p=0.06$). We found no significant bias among the third group of proteins ($p=0.2$). Within the third group we could not find any significant difference in terms of the proportion of enzymes found in proteins with symmetrical and nonsymmetrical repeats. Hence, there is no evidence for an over- or underrepresentation of symmetrical repeats among enzymes, possibly because many proteins with symmetrical repeats are also found in the form of complexes, as shown below.

Homologous proteins with different repeat copy number

If proteins with internal symmetrical repeats can work as analogues of homo-oligomers, one would expect to find variants of these proteins with different number of copies of the repeat and making complexes with accordingly different number of units. We started by testing the first hypothesis, that is, that homologues of these proteins exist with different number of copies of the repeat. For the two groups of highly repetitive proteins and α-helix-rich repeat proteins, the existence of different repeat multiplicity is largely expected by the repetitive nature of the former and by the high frequency of α-helices in the latter. We thus conservatively did the analysis on the third group of 99 proteins. Few of the sequences of naturally existing proteins are present in the sequence data banks and even fewer have a known structure. Since most of the structural repeats of this group show significant sequence similarity, we could search for homologues both at the structural level, in the PDB Cluster50 set, and at the sequence level, in the TrEMBL data bank. The use of the latter enlarged very significantly the data set size and alleviated the problem associated with the very large number of PDB structures that only cover partially the protein sequence. We searched for homologues that aligned well along all the protein, except in the repeat regions. We confirmed that many proteins have homologues with fewer copies of the repeat both among proteins with symmetrical and nonsymmetrical repeats (Table 1). This shows that functional forms of proteins exist with varying number of copies. This is consistent with our hypothesis that proteins with symmetrical repeats might be functional analogues of homo-oligomers. To better establish this point, we investigated the differences between proteins with symmetrical and nonsymmetrical repeats in terms of their functions and their quaternary structure in relation to that of proteins with fewer copies of the repeat.

Differences between symmetrical and nonsymmetrical proteins

By looking for proteins with fewer or more copies, we found some very significant differences between proteins with symmetrical and nonsymmetrical repeats. We observed very few cases of further amplifications among repeats with rotational symmetry (4 out of 47) and a significantly larger number among the others (14 out of 52, $p<0.02$, χ^2 test). This is expected from our hypothesis: such further amplifications in proteins with rotational symmetrical repeats are likely to result in disruption of the symmetry and thus be counterselected. On the other hand, proteins with nonsymmetrical repeats are expected to be less constrained in this respect.

Some proteins are found with varying numbers of repeats' copies (Table 1): 14 proteins have homologues with at least two other multiplicities (one

† http://www.ebi.ac.uk/thornton-srv/databases/pdbsum/data/pdb_EC

Table 1. Number of homologous proteins with lower or higher multiplicity in sequence database and in the Cluster50 PDB subset

		Homologous proteins			
	Query total	Different multiplicity[a]	Lower multiplicity[b]	Higher multiplicity[c]	High number of multiplicity[d]
Symmetrical proteins	47	26	23	4	2
Nonsymmetrical proteins	52	31	28	14	12

Some proteins have both homologues with lower multiplicity and homologues with higher multiplicity.
[a] At least one homologous protein exists with a different number of copies of the repeat.
[b] At least one homologous protein exists with a lower number of copies of the repeat.
[c] At least one homologous protein exists with a higher number of copies of the repeat.
[d] At least two homologous proteins exist with different numbers of copies of the repeat.

copy and/or more than two copies). Among these, only 2 correspond to proteins with symmetrical repeats, which is much less than expected by chance ($p=0.0087$, Fisher exact test). About a fourth (12 out of 52) of the proteins with nonsymmetrical repeats have at least two other repeat multiplicities.

We then checked if proteins with symmetrical and nonsymmetrical repeats were of different lengths. We did not find any significant difference by looking at the length of the structures (Supplementary Fig. 1). However many PDB structures are partial and do not correspond to the entire protein. We therefore compared the length of the protein coding sequences in our TrEMBL-based sequence data bank and found that coding sequences of proteins with symmetrical repeats are significantly shorter than that of the others (see Supplementary Fig. 1, Wilcoxon one sided test, $p=2.5\times10^{-5}$). We then calculated the number of copies of the repeats in coding sequences and found that it was strictly superior to two for 10 proteins with nonsymmetrical repeats and for only 2 proteins with symmetrical repeats. In short, proteins with nonsymmetrical repeats tend to be longer and have a higher variability in repeat copy-number.

Proteins with symmetrical repeats have homologues with fewer copies and higher oligomerisation state

We then tested if proteins with one single copy of the repeat had an oligomerisation state double to that of homologous proteins with two symmetrical copies of the repeat. We searched the PQS database‡ and the primary literature for the quaternary structure of PDB structures with symmetrical repeats having a counterpart with one single copy of the repeat. We found 11 pairs of homologous proteins with known quaternary structure in which one protein contains one copy and the other contains two copies of the repeat (Supplementary Table 4). Six pairs accurately matched our predictions (see Fig. 4 for two examples). Within the five remaining pairs, three are made up of partial structures and therefore they cannot be used as clear-cut examples. In one of the two remaining pairs both proteins are monomeric. This is the only example in clear disagreement with our prediction. The protein with only one copy of the repeat (2bv2) is a betagamma-crystallin of *Ciona intestinalis*, which split from the vertebrate lineage before the evolution of the lens.[35] 2bv2 is very similar to domains of vertebrate betagamma-crystallin, but it is composed of only one monomeric domain. The vertebrate betagamma-crystallins may have evolved from a single-domain protein. In the last pair, the two repeats cover two-thirds of the overall structure (1w25)[36] and correspond to a response regulator receiver domain that is a well-conserved domain of the two-component signal transduction system. Probably these pairs of homologues have different functions and cannot be seen as functional homologues.

Among the six pairs of proteins consistent with our predictions (see Figs. 4 and 5 and Supplementary Fig. 2), one protein (1gtt) with two copies of the repeat is a monomer, whereas the protein without a repeat is a homodimer. In two other cases (1aln and 1ddz), the former is a homodimer and the latter is a homotetramer. The last three families of proteins (2gvh, 1h9m and 1knc) are homotrimers when they have repeats, and the corresponding homologues without repeats are hexamers. These results are in exact agreement with our hypothesis that symmetrical amplifications play a role in the evolution of the oligomerisation state of proteins.

Phylogenetic analysis of protein families

We analysed the evolution of the six protein families for which we found homologues with and without the symmetrical repeat, that is, proteins with two copies or one copy of the structural repeat. For each pair we searched for the presence of homologues with one or two copies of the repeat in 849 genomes (57 archaea, 746 bacteria and 46 eukaryota) (see Materials and Methods). We only used complete genomes to ensure that lack of homologues does not result from incomplete sequencing but only from absence from the genome. For two protein families (families of 1ddz and 1h9m) we could only find homologues for one type of proteins, either with or without the repeat (Fig. 6a). This means that for some of the structures, we could

‡ ftp://ftp.ebi.ac.uk/pub/databases/msd/pqs/

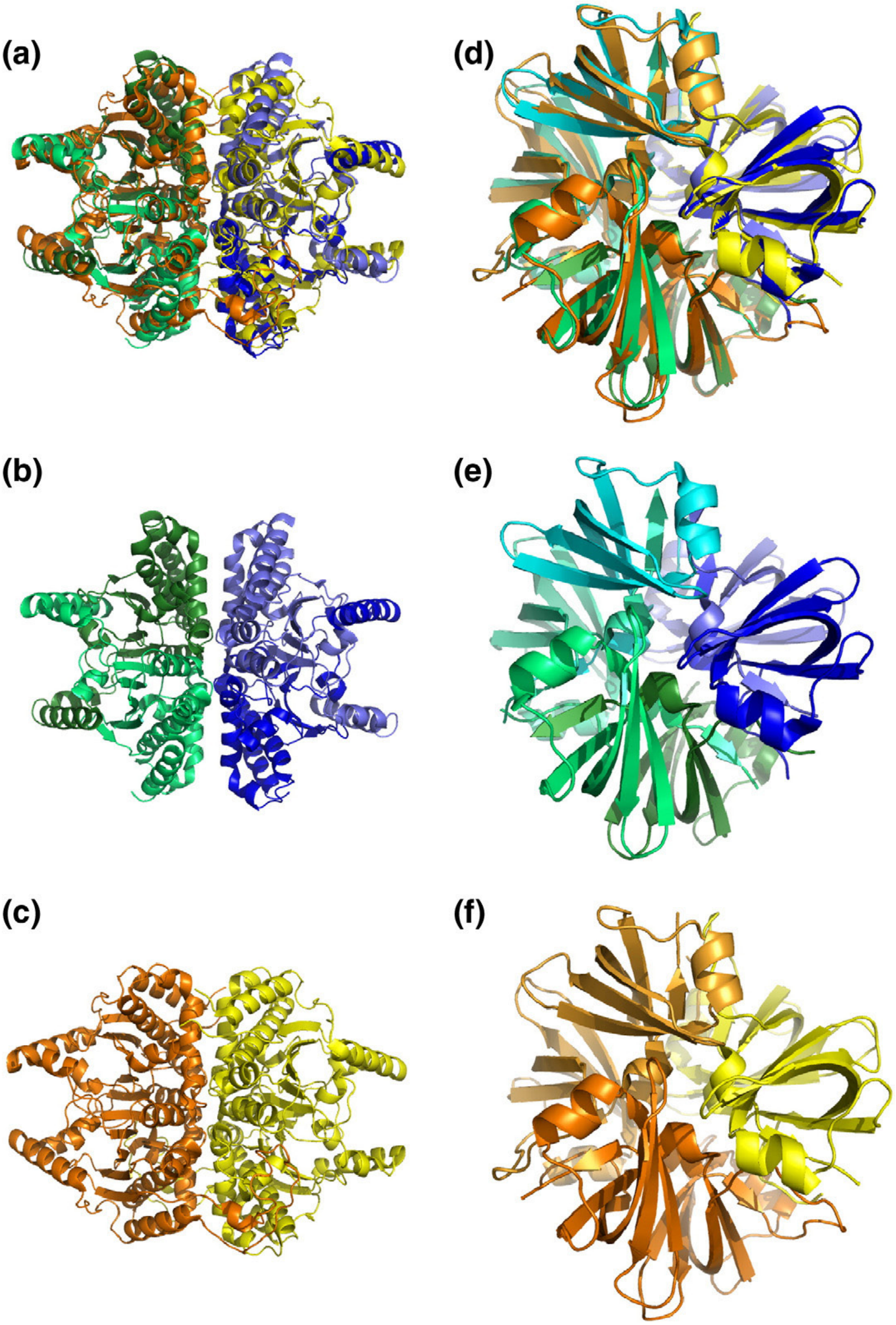

Fig. 4. Example of two proteins with a 2-fold symmetry and their homologues lacking the repeat. (a) superimposition of 1ddz and 1i6o. (b) 1i6o, β-carbonic anhydrase from *E. coli* (four chains, in blue, slate, forest and lime green). (c) 1ddz, β-carbonic anhydrase from *P. purpureum* (two chains, in yellow and orange). (d) Superimposition of 1h9m and 1fr3. (e) 1fr3, molybdate–tungstate binding protein from *S. ovata* (six chains, in blue, slate, forest, lime green, cyan and green-cyan). (f) 1h9m, molybdate-binding protein from *A. vinelandii* (three chains, in yellow, yellow-orange and orange).

not find a single occurrence of a homologous gene in completely sequenced genomes. Accordingly, the proteins lacking homologues in complete genomes have a PDB representative from unsequenced genomes, *Porphyridium purpureum* (1ddz) and *Azotobacter vinelandii* (no repeat homologues of 1h9m).

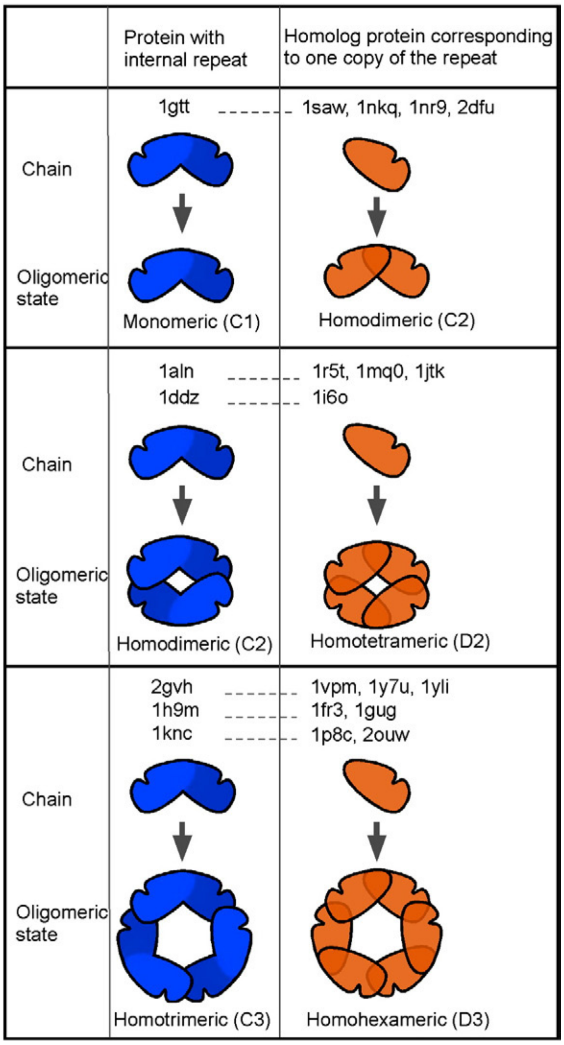

Fig. 5. Quaternary structure of pairs of homologues with two and one copy of 180° symmetrical repeats.

In three other cases (1gtt, 1aln, and 2gvh), the proteins with one single copy of the repeat are much more frequent and have a broader phylogenetic span than the proteins with two symmetrical copies of the repeat. This suggests that the former correspond to the ancestral form. In two of these cases, proteins lacking the repeat exist in archaea, bacteria and eukaryotic genomes, suggesting a very ancient origin.

Homologues of the PDB entry 1knc with or without the repeat exist in very diverse bacterial genomes. To analyse this case in more detail, we aligned the proteins of this family, removed the unaligned parts, and produced a phylogenetic tree (see Materials and Methods). We could not root the tree with a paralog, because we could not find one sufficiently similar. Therefore, we used a midpoint root, which should effectively correspond to the root if proteins evolved at similar rates in their homologous regions. This tree shows three clades with high support, suggesting that the form of the protein without the repeat is ancestral and that repeats arose later and only once in evolution (Fig. 6b). To test the frequency with which other repeats arose in evolutionary history, we repeated the phylogenetic analysis for the other three groups of proteins. One case, 1gtt, corresponds to a repeat that arose once and is present exclusively in enterobacteria. Homologues of the entry 1aln also show a monophyletic origin for the repeat. These results suggest that, in general, the proteins without the repeat arose first and that repeats tend to be monophyletic, that is, belong to one single lineage.

Yet, we found exceptions for both trends. The group of 1h9m is very surprising, as it shows that the protein with the repeat is widespread in genomes, whereas we could not find the gene for a protein without the repeat in any complete genome. This is most parsimoniously explained by claiming ancestrality of the protein with the repeat. Yet, since the two copies of the repeat share sequence similarity, they have probably arisen from an ancient duplication, which requires that a protein without the repeat preceded the extant protein with the repeat. Two scenarios could explain this apparently paradoxical result. (1) The ancestral protein without repeats disappeared after duplication, and a recent deletion created a new one-copy repeat protein that is now present in few closely related organisms. (2) The protein without repeats remained in some, but few, genomes. Since both PDB entries without repeats are from closely related Firmicutes (*Clostridium pasteurianum* and *Sporomusa ovata*), the first hypothesis seems more likely. The other unexpected case concerns 2gvh, a family where the protein with the repeat is rare among sequenced genomes, and phylogenetic evidence strongly suggests it has appeared three times independently, once within archaea (present in *Sulfolobus*), once within α-Proteobacteria (present in *Maricaulis maris*), and once in Actinobacteria (present in *Arthrobacter*). Note that this cannot be explained by lateral transfer, because the phylogenetic trees in Fig. 6b are for the protein itself, not for the species phylogeny. Thus, the proteins from these three clades are effectively very different and cluster with proteins lacking the repeat. Hence, while our results suggest that extant repeats generally arose from ancestral proteins lacking repeats, this may not always be the case. Homologous proteins with repeats may have multiple origins and they may have given rise to proteins that lost one of the copies of the repeat.

Discussion

We have shown that long internal structural repeats often show symmetry along a rotation axis. We have also shown that when proteins with symmetrical repeats have homologues lacking repeats and both have a known quaternary structure, the latter have a correspondingly higher homo-oligomerisation state. This is in agreement with our hypothesis that internal duplications may lead to the

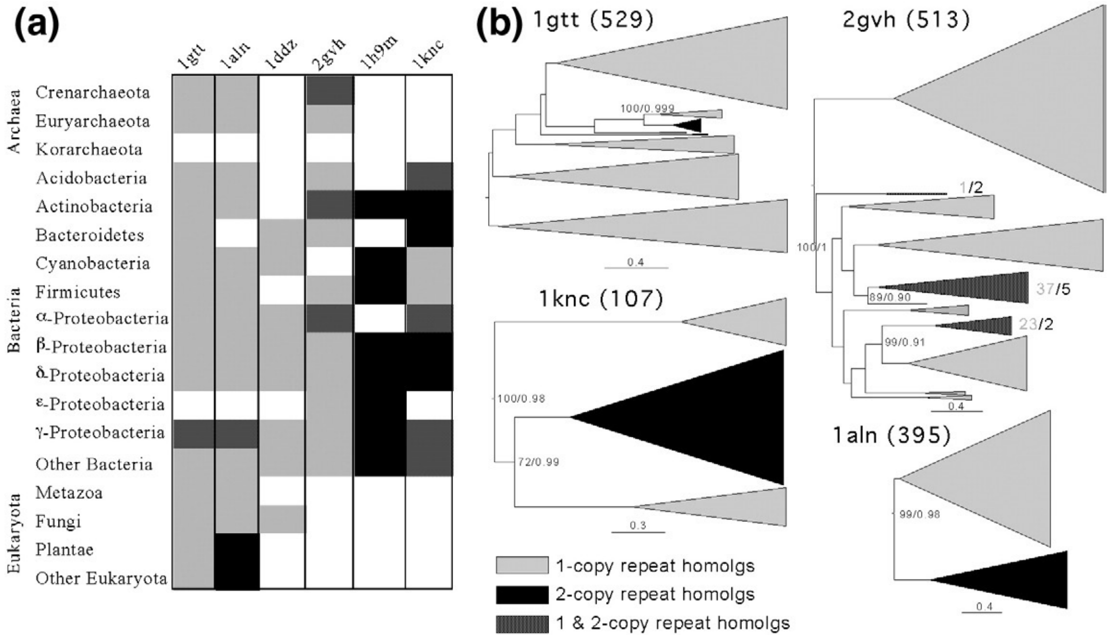

Fig. 6. Evolutionary analysis of the six families of proteins with symmetrical repeats and homologues lacking them. (a) Presence of homologues in major clades of the tree of life. (b) Cartoons of the phylogenetic trees of the four families having homologues of repeat-containing and repeat-lacking proteins in genomes (see Materials and Methods). Numbers between parentheses in the header indicate the number of proteins used to make the tree. The height of the triangles is proportional to the number of sequences in the clade. The numbers in the nodes are aLRT statistics and bootstrap values (out of 100) and the numbers in the leaves represent (for 2gvh) the number of proteins in the clade with or without the symmetrical repeat.

replacement of an n homo-oligomer by an $n/2$ homo-oligomer preserving the overall structure. We also show that among the two groups, the protein without repeats tends to be the ancestral proteins, in agreement with the idea that such repeats arise from internal genetic amplifications from preexisting proteins. Furthermore, we found that in most cases, the homo-oligomerisation state of proteins with symmetrical repeats is higher than 1, that is, such proteins still form oligomers, and typically homo-oligomers, but with fewer elements. This means that some purposed advantages of homo-oligomers, such as facilitation of allostery, may still apply. Overall, these results suggest that such proteins arise from preexisting proteins that already folded as homo-oligomers. The alternative would be a genetic amplification leading to a symmetrical repeat within a functional protein concomitantly generating *de novo* a structure resembling a homo-oligomer. This seems less likely, especially in the light of our phylogenetic analysis. We thus favour a scenario where symmetrical repeats arise in proteins that have previously evolved toward the capacity of establishing homo-oligomers. Such amplifications decrease the oligomerisation number, but do not require the creation of interfaces *de novo*, since an analogous protein complex preexisted thanks to the assembly of a larger number of monomers. Interfaces between the $n/2$ homomers remain the same as their corresponding parts in the n homomers. If such amplifications lead to a protein capable of folding correctly and performing the same function, then the fixation of this change in populations may occur by purely neutral mechanisms. Naturally, such fixation will be much more likely, and quicker, if it leads to increased fitness. Since genetic amplifications occur at high rates in genomes, the process may occur several times in parallel. This may explain the multiple emergence of symmetrical repeats in the family of acyl-CoA hydrolase (2gvh), as shown in Fig. 6. Yet, since only one out of six families showed occurrence of multiple independent genetic amplifications leading to symmetrical repeats, it is possible that this constitutes a relatively rare event, or that most such events are deleterious and are removed by natural selection.

Following these results, several questions come immediately to mind: Why are these repeats symmetrical? What are the consequences of their appearance? Are there advantages in diminishing oligomerisation number at the cost of increasing protein size?

As a reviewer pointed out to us, there is probably no selective advantage in symmetry. Yet, around half of the large internal repeats are symmetrical. This seems much too large to be due to chance. One could explain the existence of repeat symmetry in at least two ways: because symmetry is associated with selective features or because of evolutionary contingency. Symmetry could be intrinsically advantageous because of lower-energy interactions for symmetrical complexes *versus* nonsymmetrical ones.[5] The second possibility results directly from the model described in the previous paragraph:

Repeats arising within homo-oligomers will not disrupt protein function if, among other constrains, they have little impact on protein structure. If amplifications are large, they can be facilitated if they lead to a protein that has a structure close to that of the former homo-oligomer. The large overrepresentation of repeats with C2 symmetry suggests that in general, this process involves one very large amplification, leading to a protein resembling the association of two monomers of the homologous protein lacking the repeat. Such amplification will then often be symmetrical just because homo-oligomers tend to be symmetrical. If historical contingency was the only reason leading to symmetry, one would expect that such symmetry could evolve and eventually disappear. Yet, that might take a long time, since protein structures evolve slowly. Such a process has been proposed for the evolution of single-stranded DNA binding (ssb) proteins. While in most bacteria ssb protein has one OB-fold and folds as a tetramer, the ssb protein of *Deinococcus radiodurans* is a dimer and has a nearly symmetrical large repeat doubling its OB-folds.[37] This has been proposed to contribute to its extraordinary radio-resistance. Thus, in this case, an internal amplification led to a nearly symmetrical repeat that folds into a complex with half the monomers. The small deviation from a 180° rotation axis might have resulted from the amplification itself or from the relaxed structural evolution from symmetry allowed by the presence of the two OB-folds in the same peptide.

We analysed the literature to understand the consequences of the creation of symmetrical repeats in the six protein families having homologues with and without repeats and for which quaternary structure is known. Some studies on these proteins have previously noted the existence of symmetrical repeats and found or proposed functional consequences for them. For example, it has been suggested that cytidine deaminase (1aln) arose by duplication.[38] The ensuing evolution led to the loss of its zinc-coordinating residues and thereby its catalytic activity at the C-terminal domain. Indeed, the tetrameric protein has four active sites (one per monomer) compared to two active sites for the dimeric protein. Interestingly, while we found the protein containing the symmetrical repeat in 200 genomes and the protein lacking the repeat in 118 genomes, only 2 genomes contain both (*Pseudoalteromonas haloplanktis* and *Serratia proteamaculans*). This suggests that the evolution of the proteins with and without the repeat has not obliterated their functional redundancy.

Is there a specific advantage for a symmetrical repeat over an equivalent homo-oligomer? Internal symmetrical repeats have an element of symmetry that leaves the rest of the protein free to evolve away from the constraints of perfect symmetry. This slight asymmetry can be structural as in the abovementioned case of the *D. radiodurans* ssb protein, but it can also be functional. For example, the internal amplification in HpcE (an isomerase, 1gtt) allowed the evolution of a novel active site for a complementary reaction.[39] The protein consists of two very similar domains, one catalyzing a decarboxylation and the other an isomerization; these two reactions are two consecutive steps in the breaking down of homoprotocatechuate, an aromatic compound. The active site is found at the same place in the two domains. Hence, in this case, the internal amplification led to the creation of a new catalytic capability that allows the same protein to accomplish two consecutive steps in a pathway. There are other routes to create asymmetry in homo-oligomers, but they involve full gene duplication followed by divergence and adaptation.[40] The relative frequency of these different scenarios and how likely they are of producing a functionally relevant proteins is unclear. One may speculate on other advantages of larger proteins with symmetrical repeats over smaller oligomerising proteins: (1) Oligomers are unstable at low concentrations. (2) Misfolded homo-oligomers may result in protein aggregation,[41] which originates more than 30 different pathological states in humans. Accordingly, homo-oligomers endure stronger selection against the propensity to aggregate than the other proteins.[42] (4) Homo-oligomers are wholly symmetrical structures,[2] which could render them less robust to mutations. Inversely, homo-oligomers probably have several advantages over proteins with internal symmetrical repeats, as they allow modularity and they may be easier to create if indeed symmetrical repeats arise from proteins that are already homo-oligomers. Further work will be necessary to ascertain these speculations.

Homo-oligomerisation and internal repeats may also have complementary roles. This could explain why we found homo-oligomers with monomers containing internal symmetrical repeats. While the current structural PDB has been shown to strongly underrepresent proteins containing repeats,[43] the increase of available structural data will allow testing of the different possible evolutionary scenarios, leading to the selection of internal symmetrical amplifications or homo-oligomers. It will also allow quantification of the predictive potential of our observations. If a protein has a symmetrical repeat, it will be important to know the probability that a homologue lacking the repeat folds symmetrically as a homo-oligomer. Conversely, our work shows that homologues of homo-oligomers with large symmetrical repeats will tend to have lower oligomerisation numbers.

Materials and Methods

Identification of long repeats

We used Swelfe[22] to find internal repeats in protein structures of the Astral compendium.[23] Parameters were adapted to find long repeats (length of repeat >50 α angles, score >200, gap open penalty = 70, gap extension penalty = 30, relative RMSD <0.5). We kept only the highest-scoring repeat in each structure. Out of the 188 proteins containing such long repeats, we suppressed all with more than 40% sequence similarity with other proteins. This left 172 proteins for further analysis. Using HMMER2, we identified Pfam domains in the proteins.[44]

Symmetrical long repeats

To assess the symmetry of the repeats, we computed the angle allowing best superimposition of the two copies of the repeat in the protein structure with the Zuker and Somorjai program.[45] Repeats obtained by a rotation of an angle higher than 170° or lower than −170° correspond to a 2-fold symmetry. We considered an interval of −5°/+5° around ±60°, ±90°, ±120° to compute the number of proteins with repeats showing 6-, 4- and 3-fold symmetry, respectively.

Looking for one-copy or *n*-copy proteins

We searched for proteins with varying numbers of repeat copies both in structures and in amino acid sequences.

Three-dimensional structures

We searched for similar elements of the structural repeat in the nonredundant data bank Cluster50 from the PDB[46] (less than 50% sequence identity) with Swelfe (with the option for comparing two sequences or two structures). We did not use Astral in this analysis because it filters redundant domains and thus automatically eliminates the proteins with varying numbers of repeats. Proteins with elements matching the structural repeats were then selected according to their length: it must differ from the query protein from at least 0.5 times the length of the repeat. Prospective proteins were then superimposed with the initial proteins with Pymol[47] to check that they contain different numbers of the repeat.

Finding full amino acid sequences corresponding to PDB entries

The amino acid sequence corresponding to the structural repeat was compared with all proteins in UniProtKB/TrEMBL[48] (Supplementary Fig. 3). A first rapid comparison was made with BlastP[49] (with default parameters) to retrieve candidate proteins from the data bank. Then we used Swelfe to find proteins containing at least one copy of the repeat. The length of the matched copies must be at least three-fourths that of the reference copy. After that, an end-gap free global alignment (using the BLOSUM62 matrix, gap opening 1.2, gap extension 0.8) was achieved between complete sequences with repeats and proteins having at least one copy of the repeat. We kept the proteins leading to alignments with >40% similarity. For one-copy proteins, we removed proteins longer than the original protein minus half of the length of the repeat. We then checked that positions corresponding to one copy of the repeat were aligned (less than 20% of gaps) and that positions corresponding to the other copy matched with gaps (more than 80% of gaps).

To find homologous proteins with higher number of copies of the repeat, we kept proteins longer than the sum of the length of query protein and 0.5 times the repeat length. Then, using Swelfe, we checked that these proteins contain more copies of the repeat than the original protein. The length of the matched copies must be at least three quarters of that of the query repeat. When homologous proteins with the same number of copies of the repeat were found, only one was kept.

Phylogenetic analysis

Homologous sequences with one or two copies of the repeat were searched with BLAST[49] against 849 complete proteomes (57 archaea, 746 bacteria and 46 eukaryota). Hits with e-value <1e-10 were conserved. The length of homologous sequences differed by less than 20% from that of the query sequence. The sets of homologues with one- and two-copy repeats were put together and aligned with MUSCLE with default parameters.[50] The multiple alignments were then inspected visually with SEAVIEW[51] to remove regions that were poorly aligned or exclusive to two-copy repeats. The multiple alignments were then used as input to PHYML[52] to make the phylogenetic trees using maximum likelihood with the WAG matrix and a gamma correction for variable evolutionary rates. Robustness of the branches was tested with the aLRT test of PHYML and with nonparametric bootstrap using Puzzle.[53] Trees were drawn using FigTree§.

Acknowledgements

A.-L.A. acknowledges support from the Conseil Régional d'Ile de France. This project was partly supported by the PROTEUS project from ANR-06-CIS (Calcul Intensif et Simulation) and ACI EVOLREP. We thank three anonymous reviewers for comments, criticisms and suggestions that significantly improved this manuscript. We thank Etienne Loire for the program to compute biases in GO terms.

Supplementary Data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jmb.2009.09.031

References

1. Goodsell, D. S. & Olson, A. J. (2000). Structural symmetry and protein function. *Annu. Rev. Biophys. Biomol. Struct.* **29**, 105–153.
2. Monod, J., Wyman, J. & Changeux, J. P. (1965). On the nature of allosteric transitions: a plausible model. *J. Mol. Biol.* **12**, 88–118.
3. Blundell, T. L. & Srinivasan, N. (1996). Symmetry, stability, and dynamics of multidomain and multi-component protein systems. *Proc. Natl Acad. Sci. USA*, **93**, 14243–14248.
4. Wolynes, P. G. (1996). Symmetry and the energy landscapes of biomolecules. *Proc. Natl Acad. Sci. USA*, **93**, 14249–14255.

§ http://tree.bio.ed.ac.uk/software/figtree/

5. Andre, I., Strauss, C. E., Kaplan, D. B., Bradley, P. & Baker, D. (2008). Emergence of symmetry in homooligomeric biological assemblies. *Proc. Natl Acad. Sci. USA*, **105**, 16148–16152.
6. Levy, E. D., Boeri Erba, E., Robinson, C. V. & Teichmann, S. A. (2008). Assembly reflects evolution of protein complexes. *Nature*, **453**, 1262–1265.
7. Levy, E. D., Pereira-Leal, J. B., Chothia, C. & Teichmann, S. A. (2006). 3D complex: a structural classification of protein complexes. *PLoS Comput. Biol.* **2**, e155.
8. Pereira-Leal, J. B., Levy, E. D. & Teichmann, S. A. (2006). The origins and evolution of functional modules: lessons from protein complexes. *Philos. Trans. R. Soc. London, Ser. B*, **361**, 507–517.
9. Goodsell, D. S. & Olson, A. J. (1993). Soluble proteins: size, shape and function. *Trends Biochem. Sci.* **18**, 65–68.
10. Bucciantini, M., Giannoni, E., Chiti, F., Baroni, F., Formigli, L., Zurdo, J. *et al.* (2002). Inherent toxicity of aggregates implies a common mechanism for protein misfolding diseases. *Nature*, **416**, 507–511.
11. Kirschner, M. & Gerhart, J. (1998). Evolvability. *Proc. Natl. Acad. Sci. USA*, **95**, 8420–8427.
12. Marcotte, E. M., Pellegrini, M., Yeates, T. O. & Eisenberg, D. (1999). A census of protein repeats. *J. Mol. Biol.* **293**, 151–160.
13. Rocha, E. P. (2003). An appraisal of the potential for illegitimate recombination in bacterial genomes and its consequences: from duplications to genome reduction. *Genome Res.* **13**, 1123–1132.
14. Bjorklund, A. K., Ekman, D. & Elofsson, A. (2006). Expansion of protein domain repeats. *PLoS Comput. Biol.* **2**, e114.
15. Treangen, T. J., Abraham, A. L., Touchon, M. & Rocha, E. P. (2009). Genesis, effects and fates of repeats in prokaryotic genomes. *FEMS Microbiol. Rev.* **33**, 539–571.
16. Lavorgna, G., Patthy, L. & Boncinelli, E. (2001). Were protein internal repeats formed by "bricolage"? *Trends Genet.* **17**, 120–123.
17. Apic, G., Gough, J. & Teichmann, S. A. (2001). An insight into domain combinations. *Bioinformatics*, **17** (Suppl. 1), S83–S89.
18. Andrade, M. A., Perez-Iratxeta, C. & Ponting, C. P. (2001). Protein repeats: structures, functions, and evolution. *J. Struct. Biol.* **134**, 117–131.
19. Pereira-Leal, J. B. & Teichmann, S. A. (2005). Novel specificities emerge by stepwise duplication of functional modules. *Genome Res.* **15**, 552–559.
20. Bjorklund, A. K., Ekman, D., Light, S., Frey-Skott, J. & Elofsson, A. (2005). Domain rearrangements in protein evolution. *J. Mol. Biol.* **353**, 911–923.
21. Pasek, S., Risler, J. L. & Brezellec, P. (2006). Gene fusion/fission is a major contributor to evolution of multi-domain bacterial proteins. *Bioinformatics*, **22**, 1418–1423.
22. Abraham, A. L., Rocha, E. P. & Pothier, J. (2008). Swelfe: a detector of internal repeats in sequences and structures. *Bioinformatics*, **24**, 1536–1537.
23. Chandonia, J. M., Hon, G., Walker, N. S., Lo Conte, L., Koehl, P., Levitt, M. & Brenner, S. E. (2004). The ASTRAL Compendium in 2004. *Nucleic Acids Res.* **32**, D189–D192.
24. Betancourt, M. R. & Skolnick, J. (2001). Universal similarity measure for comparing protein structures. *Biopolymers*, **59**, 305–309.
25. Cheng, H., Kim, B. H. & Grishin, N. V. (2008). Discrimination between distant homologs and structural analogs: lessons from manually constructed, reliable data sets. *J. Mol. Biol.* **377**, 1265–1278.
26. Wiener, M., Freymann, D., Ghosh, P. & Stroud, R. M. (1997). Crystal structure of colicin Ia. *Nature*, **385**, 461–464.
27. Eswaramoorthy, S., Kumaran, D., Keller, J. & Swaminathan, S. (2004). Role of metals in the biological activity of Clostridium botulinum neurotoxins. *Biochemistry*, **43**, 2209–2216.
28. Macedo, S., Romao, C. V., Mitchell, E., Matias, P. M., Liu, M. Y., Xavier, A. V. *et al.* (2003). The nature of the di-iron site in the bacterioferritin from *Desulfovibrio desulfuricans*. *Nat. Struct. Biol.* **10**, 285–290.
29. Pickles, L. M., Roe, S. M., Hemingway, E. J., Stifani, S. & Pearl, L. H. (2002). Crystal structure of the C-terminal WD40 repeat domain of the human Groucho/TLE1 transcriptional corepressor. *Structure*, **10**, 751–761.
30. Jing, H., Takagi, J., Liu, J. H., Lindgren, S., Zhang, R. G., Joachimiak, A. *et al.* (2002). Archaeal surface layer proteins contain beta propeller, PKD, and beta helix domains and are related to metazoan cell surface proteins. *Structure*, **10**, 1453–1464.
31. Krieger, I., Kostyukova, A., Yamashita, A., Nitanai, Y. & Maeda, Y. (2002). Crystal structure of the C-terminal half of tropomodulin and structural basis of actin filament pointed-end capping. *Biophys. J.* **83**, 2716–2725.
32. Michaely, P., Tomchick, D. R., Machius, M. & Anderson, R. G. (2002). Crystal structure of a 12 ANK repeat stack from human ankyrinR. *EMBO J.* **21**, 6387–6396.
33. Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M. *et al.* (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25**, 25–29.
34. Loire, E., Praz, F., Higuet, D., Netter, P. & Achaz, G. (2009). Hypermutability of genes in *Homo sapiens* due to the hosting of long mono-SSR. *Mol. Biol. Evol.* **26**, 111–121.
35. Shimeld, S. M., Purkiss, A. G., Dirks, R. P., Bateman, O. A., Slingsby, C. & Lubsen, N. H. (2005). Urochordate betagamma-crystallin and the evolutionary origin of the vertebrate eye lens. *Curr. Biol.* **15**, 1684–1689.
36. Chan, C., Paul, R., Samoray, D., Amiot, N. C., Giese, B., Jenal, U. & Schirmer, T. (2004). Structural basis of activity and allosteric control of diguanylate cyclase. *Proc. Natl Acad. Sci. USA*, **101**, 17084–17089.
37. Bernstein, D. A., Eggington, J. M., Killoran, M. P., Misic, A. M., Cox, M. M. & Keck, J. L. (2004). Crystal structure of the *Deinococcus radiodurans* single-stranded DNA-binding protein suggests a mechanism for coping with DNA damage. *Proc. Natl Acad. Sci. USA*, **101**, 8575–8580.
38. Xie, K., Sowden, M. P., Dance, G. S., Torelli, A. T., Smith, H. C. & Wedekind, J. E. (2004). The structure of a yeast RNA-editing deaminase provides insight into the fold and function of activation-induced deaminase and APOBEC-1. *Proc. Natl Acad. Sci. USA*, **101**, 8114–8119.
39. Tame, J. R., Namba, K., Dodson, E. J. & Roper, D. I. (2002). The crystal structure of HpcE, a bifunctional decarboxylase/isomerase with a multifunctional fold. *Biochemistry*, **41**, 2982–2989.
40. Pereira-Leal, J. B., Levy, E. D., Kamp, C. & Teichmann, S. A. (2007). Evolution of protein complexes by duplication of homomeric interactions. *Genome Biol.* **8**, R51.

41. Ding, F., Dokholyan, N. V., Buldyrev, S. V., Stanley, H. E. & Shakhnovich, E. I. (2002). Molecular dynamics simulation of the SH3 domain aggregation suggests a generic amyloidogenesis mechanism. *J. Mol. Biol.* **324**, 851–857.
42. Chen, Y. & Dokholyan, N. V. (2008). Natural selection against protein aggregation on self-interacting and essential proteins in yeast, fly, and worm. *Mol. Biol. Evol.* **25**, 1530–1533.
43. Peng, K., Obradovic, Z. & Vucetic, S. (2004). Exploring bias in the Protein Data Bank using contrast classifiers. *Pac. Symp. Biocomput.*, 435–446.
44. Finn, R. D., Tate, J., Mistry, J., Coggill, P. C., Sammut, S. J., Hotz, H. R. *et al.* (2008). The Pfam protein families database. *Nucleic Acids Res.* **36**, D281–288.
45. Zuker, M. & Somorjai, R. L. (1989). The alignment of protein structures in three dimensions. *Bull. Math. Biol.* **51**, 55–78.
46. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H. *et al.* (2000). The Protein Data Bank. *Nucleic Acids Res.* **28**, 235–242.
47. Ordog, R. (2008). PyDeT, a PyMOL plug-in for visualizing geometric concepts around proteins. *Bioinformation*, **2**, 346–347.
48. Wu, C. H., Apweiler, R., Bairoch, A., Natale, D. A., Barker, W. C., Boeckmann, B. *et al.* (2006). The Universal Protein Resource (UniProt): an expanding universe of protein information. *Nucleic Acids Res.* **34**, D187–D191.
49. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids. Res.* **25**, 3389–3402.
50. Edgar, R. C. (2004). MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, **5**, 113.
51. Galtier, N., Gouy, M. & Gautier, C. (1996). SEAVIEW and PHYLO_WIN: two graphic tools for sequence alignment and molecular phylogeny. *Comput. Appl. Biosci.* **12**, 543–548.
52. Guindon, S. & Gascuel, O. (2003). A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst. Biol.* **52**, 696–704.
53. Schmidt, H. A., Strimmer, K., Vingron, M. & von Haeseler, A. (2002). TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics*, **18**, 502–504.