Published in final edited form as:
J Mol Biol. 2004 May 7; 338(4): 633–641. doi:10.1016/j.jmb.2004.03.039.

Domain Insertions in Protein Structures

R. Aroul-Selvam[1], **Tim Hubbard**[1], and **Rajkumar Sasidharan**[2,*]

[1]The Wellcome Trust Sanger Institute, Genome Campus Hinxton, Cambridge CB10 1SA UK

[2]MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, UK

Abstract

Domains are the structural, functional or evolutionary units of proteins. Proteins can comprise a single domain or a combination of domains. In multi-domain proteins, the domains almost always occur end-to-end, i.e., one domain follows the C-terminal end of another domain. However, there are exceptions to this common pattern, where multi-domain proteins are formed by insertion of one domain (insert) into another domain (parent). Here, we provide a quantitative description of known insertions in the Protein Data Bank (PDB). We found that 9% of domain combinations observed in non-redundant PDB are insertions. Although 90% of all insertions involve only one insert, proteins can clearly have multiple (nested, two-domain and three-domain) inserts. We also observed correlations between the structure and function of a domain and its tendency to be found as a parent or an insert. There is a bias in insert position towards the C terminus of parents. We observed that the atomic distance between the N and C terminus of an insert is significantly smaller when compared to the N-to-C distance in a parent context or a single domain context. Insertions are found always to occur in loop regions of parent domains. Our observations regarding the relationship between domain insertions and the structure, function and evolution of proteins have implications for protein engineering.

Keywords

domain insertion; inserted domain; discontinuous domains; non-contiguous domains; protein engineering

Introduction

It is now widely accepted that domains constitute the basic structural, functional or evolutionary unit of proteins.1-3 Proteins can comprise a single domain or they can be made from several domains resulting in a multi-domain protein. The exponential growth of protein sequence and structure data and the development of sensitive sequence comparison methods have contributed significantly towards understanding the mechanisms of protein evolution. Sequence and structure-based comparison of protein database sequences suggested that evolution made use of a limited repertoire of domain families to create multi-domain proteins with a wide variety of architecture to cater to the functional requirements of an organism at the molecular level.4,5 Structural assignments to gene sequences from complete genomes revealed that about two-thirds of prokaryotic proteins and 80% of eukaryotic proteins are multi-domain proteins.6 The preponderance of multi-domain proteins in the

*Corresponding author E-mail address of the corresponding author: sraj@mrc-lmb.cam.ac.uk.

Supplementary data associated with this article can be found at doi: 10.1016/j.jmb.2004.03.039

Europe PMC Funders Author Manuscripts

three kingdoms of life underscores their role in the evolution of diverse molecular functions. Thus, it becomes important to understand the evolution of multi-domain proteins.

Domain organisation in multi-domain proteins

In 1973, Donald Wetlaufer introduced the concept of continuous and discontinuous domains.7 A continuous domain is formed by one part of a polypeptide chain, while a discontinuous domain is formed by two or more parts of a single polypeptide chain. A majority of multi-domain proteins are formed by continuous domains, where the individual domains are secured by end-to-end linkages. However, there are exceptions to this common pattern where proteins exhibit discontinuity in their domain arrangement. Here, we focus on insertions, where a domain is inserted into another domain (Figure 1). Essentially, insertions represent one example of non-contiguous domain arrangement. While domain insertions were described anecdotally in a few protein structures by Russell,8 the availability of an accurate and well-curated domain classification resource such as the Structural Classification of Proteins (SCOP) database and an ever increasing size of the Protein Data Bank (PDB) gave us an opportunity to investigate the phenomenon comprehensively. Here, we provide a quantitative description of domain insertions in 3D structures.

The SCOP database

We followed the definition of protein domains in the SCOP database (version 1.61).1 Although there are several available schemes of protein structure classification, we chose SCOP because it is an expert curated classification of protein structures based on their structural and evolutionary relatedness. In the SCOP database, a protein domain is considered as a unit of evolution if it occurs independently by itself or in combination with other domains.

SCOP represents a hierarchical classification scheme with four principal levels: family, superfamily, fold and class. Domains clustered into families are related evolutionarily and can be detected at the sequence level. Domains grouped within superfamilies can have low sequence identity, but their structural and functional features suggest a common evolutionary origin. Superfamilies with similar topology are grouped under a fold. Folds are assigned to classes based on their secondary structure. For our analysis, we considered the fold and superfamily levels of the SCOP hierarchy, and the five major classes (all-α, all-β, α/β, α + β and “small proteins”). All-α and all-β classes include proteins with abundant α-helices or β-sheets, respectively. The α/β class is distinguished mainly by parallel β-sheets (β-α-β units), whereas the α + β class contains proteins with predominantly anti-parallel β-sheets (segregated α and β regions). “Small proteins” are distinguished by their size rather than other features.

Identification of insertions

We obtained data for our analysis from the PDB.9 To overcome the redundancy inherent in the PDB, we chose a pre-computed list of non-redundant protein chains provided by PDB_Select[†].10 We used the set of proteins that had pairwise sequence identities less than 90%. We designated this set as PDB_90. Out of the 6182 chains in PDB_90, only 5883 chains were assigned SCOP domain definitions. We used the SCOP parseable file “dir.cla.scop.txt_1.61”[‡] to extract domain definitions.

[†]April 2002 release obtained from ftp://ftp.embl-heidelberg.de/pub/databases/protein_extras/pdb_select
[‡]http://scop.mrc-lmb.cam.ac.uk/scop/parse/dir.cla.scop.txt_1.61

It is self-evident that insertions can only be found in multi-domain proteins, where one domain (insert) is contained within another domain (parent). Parent and insert domains can belong to the same or different SCOP superfamilies. Likewise, a combination of two domains can be viewed as a combination of superfamily participations. We obtained a total of 140 protein chains that conformed to our definition. When we considered 140 protein chains as parent-insert superfamily participations, we observed several identical parent-insert superfamily participations. Whenever there was also the same topological relationship between the parent and insert domains, we retained only one example of a parent-insert superfamily participation. This procedure left us with 40 unique parent-insert superfamily participations. Variations on the simple scheme “one insert within one parent” are present; they are shown in Figure 2.

For all cases of identified domain insertions, we checked for artefacts arising from missing coordinates. This was necessary because SCOP domain definitions are based on atomic coordinates provided in the PDB. To ascertain consistency, we compared atomic coordinates (ATOM records) *versus* sequences (SEQRES records) obtained from the ASTRAL compendium.11 In the majority of cases, the sequences are completely covered by coordinates, but in other cases, there are parts of sequences with missing coordinates. However, the coordinates that are absent do not obscure the position of insertion in the latter cases.

We then calculated unique superfamily participations for all multi-domain proteins. We identified 450 unique superfamily participations for 5883 single or multi-domain proteins in SCOP. Thus, domain insertions constitute 9% (40/450) of all unique superfamily combinations.

Types of domain insertions

Domain insertions can be categorised as either single or multiple depending on the number of inserts (Figure 2). In single insertions, one domain is inserted into another domain, and both domains can belong to the same or different superfamilies. For example, in Figure 1, the *Escherichia coli* enzyme RNA 3′-terminal phosphate cyclase (PDB 1qmhA)12 has two domains, a small insert and a larger parent that belong to different superfamilies. 90% (36/40) of the observed insertions are single domain insertions. In multiple insertions, more than one domain, either of the same or different superfamily, is inserted into the parent domain. We observed three types of multiple insertions: (i) Nested insertions: In *Thermoplasma acidophilum* thermosome (PDB 1a6dA),13 the apical domain of the archael chaperonin is inserted into the intermediate domain, which is in turn inserted into an ATPase domain. (ii) Two-domain insertions: The type II inosine monophosphate dehydrogenase from *Streptococcus pyogenes* (PDB 1zfjA)14 contains two tandem cystathionine-β-synthase domains inserted into the catalytic TIM-barrel domain. The second example of this is the *Saccharomyces cerevisiae* PI-*Sce* I intein (PDB 1ef0A),15 a homing endonuclease with protein splicing activity, which has the duplicated endonuclease domain inserted into the Hint domain. (iii) Three-domain insertions: In PI-*Pfu* I, an intein-encoded homing endonuclease from the archaebacteria *Pyrococcus furiosus* (PDB 1dq3A),16 the Hint domain has three tandem inserts, two intein endonuclease domains with αββαββαα structural motifs, and one Stirrup domain.

Previous work on intron-encoded homing endonucleases from the dodecapeptide family showed that for their folding, dimerisation and catalysis, they should form a dimer that has two copies of the LAGLIDADG motif (one copy per subunit of a dimer), or alternatively they could be monomeric with the monomer having both copies of the motif.17 We found that in PI-*Sce* I (case (ii)) and PI-*Pfu* I (case (iii)), two monomeric domains are tandemly

inserted into one parent domain. This observation suggests to us that during the course of evolution, there was a simultaneous insertion of two monomeric domains into the parent domain, rather than an insertion of one monomeric domain followed by its duplication.

In our analysis, we treated multiple insertions as several separate parent-insert combinations, resulting in the total of 45 such combinations within 40 protein chains. There are 41 unique parent-insert superfamily combinations. Upon examination of relationships among proteins containing insertions, levels of SCOP hierarchy, and superfamily participation of parent and inserted domains, we identified several biologically meaningful patterns. These findings are discussed below.

Nature and characteristics of domain insertions: class level

As mentioned before, we considered five SCOP classes. There is a maximum of 25 (5 × 5) different class pairwise combinations. In our data, we observed only 15 combinations when investigating class participation of parent-insert pairs. The combination of α/β-parent-α + β-insert is predominant, where 50% of all parents belonged to α/β class and 40% of all inserts belonged to α + β class. Domains from α/β class occur as parent domains twice and four times more often than domains from all-β and all-α class, respectively. Domains from the class of “small proteins” are seen only as inserts. This bias could be explained, at least to a certain extent, by taking into consideration the size and function of parents and inserts, which is articulated in the next section.

Size and function of domains involved in insertions

Figure 3(a) shows the domain length distribution for proteins from PDB_90 across the five SCOP classes. The average domain length is longest for α/β class followed by the all-β, α + β, and all-α class. When we calculated distribution of average domain lengths for 41 parent domains, we observed the same trend (Figure 3(b)). However, the average length of parent domains is noticeably larger than the average length of domains from PDB_90 set; this is true for each SCOP class (compare Figure 3(a) and (b)). Thus, combining the fact that α/β parent domains are the most abundant, with the fact that α/β domains are the longest on average, we arrived at an explanation that longer domains more readily accept insertions during evolution. As for the inserted domains, α + β and all-α class are equal and major contributors to the number of domains. Therefore, the trend observed for parents is not applicable for inserts.

In most cases, inserted domains are shorter than parent domains (Figure 4(a)). Parents comprised 50–80% of protein length, while inserts comprised 20–50%. Close to 80% of inserts are shorter than 175 residues, which is the average length of a protein domain calculated from crystal structures.18 More than 60% of inserts are shorter than 130 residues. This observation is consistent with the heuristic thinking that smaller domains are less likely to disturb the structure and folding of parent domains; the observation could explain shorter lengths of inserted domains. Our explanation does not contradict an important experiment by Doi and colleagues.19 They were able to show that when random sequences of 120–130 amino acid residues were inserted into a surface loop region of *E. coli* RNase HI, about 10% of the clones retained >1% of the wild-type RNase HI activity.19

The large proportion of α/β class domains as parents can be correlated with their biochemical function. Previous work showed that more than half of the proteins in the PDB are enzymes, and close to one half of all enzyme families contain multi-domain proteins. Multi-domain enzymes often consist of a catalytic domain and a nucleotide-binding domain. 20 It is therefore possible to predict that domain insertions are likely to occur in enzymes. Indeed, in our dataset, 39 out of 40 parent-insert pairs conform to this prediction. The

remaining non-enzymatic protein is the bluetongue virus capsid protein vp-7, which has the central domain from all-β class inserted into the multi-helical parent domain. A genome-scale analysis of the structural features of proteins revealed that proteins with α/β-fold are frequently involved in fusion events.21 α/β-folds are also known to be associated disproportionately with enzymatic function,20 which lends further credence to the prominent role of α/β-folds in accepting insertions.

Nature and characteristics of domain insertions: fold and superfamily level

Out of 57 folds in the class of "small proteins", we found two domains with a similar fold (Rubredoxin) as inserts; both the inserted domains belong to the same superfamily. Within the α + β class, the 18 inserted domains (from 15 superfamilies) spanned 11 folds; there are 204 different folds in the α + β class (data not shown). The trend is similar for the other SCOP classes, where folds of inserted domains constitute minor fractions of known folds. In contrast to the inserts, all parent domains have different folds. Thus, we observed another distinction between parents and inserts at the fold level.

Similarly, parent superfamilies are found to be more versatile than insert superfamilies. Most insert superfamilies combine with only one parent superfamily. There are merely three out of 45 insert superfamilies that combine with two different parent superfamilies. These insert superfamilies are NAD(P)-binding Rossmann superfamily, FAD/NAD(P)-binding superfamily and C-terminal domain of FAD-linked reductases superfamily.

While most parent superfamilies combine with just one insert superfamily, there are five conspicuous exceptions. There are three parent super-families each combining two different insert superfamilies. The three parent superfamilies are Zn-dependent exopeptidases superfamily, nucleotidyl transferase superfamily, and nucleotide-binding domain superfamily. Moreover, there are two parent superfamilies each combining with three different insert superfamilies. The two parent superfamilies are P-loop containing NTP hydro-lases superfamily, and FAD/NAD(P)-binding domain superfamily.

Two further observations at the superfamily level are worth mentioning. Firstly, with one exception, all parents and inserts belong to different superfamilies: in the *E. coli* enzyme glutathione reductase (PDB 1gesB),22 both the parent and insert belong to the superfamily of FAD/NAD(P)-binding domains. Secondly, superfamilies that are popular in the parent or insert context also appear to be popular in sequential domain combinations.23 They are found combining with more than one superfamily in sequential domain order. One exception to this correlation is the superfamily of C-terminal domains of FAD-linked reductases; this superfamily is popular in the insert context, but does not tandemly combine with other superfamilies.

Point of insertion

We did not find any bias in the distribution of insertion points within 41 unique parent-insert combinations. However, we observed a significant bias in the location of the insertion point when we considered a subset of 28 parent-insert combinations, where either the parent or insert superfamily also participated in sequential combination with other superfamilies. As shown in Figure 4(b), for the 28 cases in question the insertion point occurred in the last third part of the parent domain sequence (confidence level 98%). Spatially, all 41 insertions are observed in loop regions of the 3D structure of parent domains.

Proximity of N and C termini in inserts

We wanted to determine how the insertion context affects the distance between N and C termini of an inserted domain. Distance between termini was defined as the distance between C^{α} atoms of the first and the last residue of the domain. We first calculated distances for domains that do not participate in insertions. In order to do this, we considered 1000 domains, each representative of a SCOP superfamily. We obtained sequences and coordinates for these 1000 domains from the ASTRAL compendium.11 Only 687 domain sequences are covered completely by coordinates. Using AEROSPACI scores,11 we were able to find 60 substitutes for the 313 representative domains that are not entirely covered by coordinates. Altogether, we obtained complete coordinate information for 747 domains (687 + 60). Because we confined our analysis to five major SCOP classes, we calculated distances between termini for 711 domains, as the rest do not belong to the five classes being investigated. The average distance for representative domains is 25 Å.

Calculation of distances between the termini of inserted domains was less straightforward. Domain boundaries reported in SCOP are defined manually. Therefore, we compared SCOP domain boundaries for 41 inserted domains against the domain boundaries reported in CATH database.24 In contrast to SCOP, CATH structural classification of proteins is produced automatically. However, only 28 out of 41 inserted domains were available in CATH. For the other 13, there were differences in domain classification or the corresponding proteins were absent from CATH classification. For 28 inserted domains, boundaries are identical between SCOP and CATH. The average distance between domain termini of inserted domains is 8 Å (confidence level 99%), which is two-thirds shorter than the distance between termini in normal domains.

There are two superfamilies that occur in both parent and insert context. This example allowed us to compare distances between termini for a parent and an insert from the same superfamily. In case of FAD/NAD(P)-binding domain superfamily, the distances are 30 Å and 5 Å for parent and insert, respectively. These figures are 11 Å and 8Å for NAD-binding Rossmann domain superfamily. Thus, our analysis unambiguously shows that the ends of inserted domains are significantly closer than the ends of parent domains, or domains not participating in insertions.

It is interesting to speculate how the distance between domain termini can affect stability and conformational flexibility of a protein domain. While insertion context might generally reduce conformational freedom of the domain, it can simultaneously contribute to the stability of the domain, which would in turn affect its function. One can also imagine how the close proximity of domain termini can restore protein conformational flexibility by mimicking an inter-domain link observed in sequentially ordered domains.

Conclusions

Utilising an evolutionary basis of domain classification, we described the nature and characteristics of domain insertions in known protein structures. Domain insertions represent an unusual but abundant case of multi-domain proteins. Our analysis provides several novel insights into the nature and characteristics of domain insertions: (1) 9% of multi-domain proteins contain insertions. (2) The majority of insertions are single domain insertions. We also found two-domain, three-domain, and nested insertions. (3) α/β class has a higher propensity to accept insertions. This can be correlated to the size and function of proteins within this class. (4) In most cases, parent domains are found to be longer than the inserted domains. (5) When fold and superfamily participations were considered for parents and inserts, the former are found to be more versatile than the latter, in that the parent domains

combined with different partners. (6) The point of insertion is biased towards the C terminus of parents, whenever the parent domain belongs to the superfamily that sequentially combines with other superfamilies. (7) Inserted domains tend to have juxtaposed termini compared to parent domains or domains that do not participate in insertions. Perhaps, these domains are more viable in the insert context when their termini are close in space; small size can further contribute to their viability.

Implications

Our results clearly indicate that, despite the structural and functional constraints inherent in the insertion of a domain into another, domain insertion an effective way to create multi-domain proteins. Functional hybrid proteins have been created through domain insertion in the laboratory by several groups. We cite three examples to support our observations. Betton and co-workers created hybrid proteins by inserting a penicillin-hydrolysing enzyme TEM β-lactamase (Bla) into the maltodextrin-binding protein (MalE);25 they used the permissive insertion sites identified before.26 Of the two insertions that resulted in functional hybrids, one insertion occurred in the first quarter of the MalE protein, while the other occurred in the last quarter. The parent protein (MalE) belongs to the α/β class; the distance between the termini of the inserted domain (Bla) is 5 Å, as shown by the authors. The proteins 1,3-1,4-β-glucanase from *Bacillus macerans* (wtGLU) and 1,4-β-xylanase from *Bacillus subtilis* (wtXYN) are single-domain jellyroll proteins catalysing similar enzymatic reactions; cpMAC-57 is a circularly permuted variant of wtGLU. Ay *et al.*, created a fusion protein by inserting wtXYN into cpMAC-57. The authors showed that both fold spontaneously and have enzyme activities at wild-type level. The crystal structure of the chimeric protein showed nearly ideal, native-like fold for both the domains.27 In the third example, Collinet *et al.* were successfully able to produce a chimeric protein with a two-domain insertion. They inserted the monomeric proteins dihydrofolate reductase (159 residues, belongs to α/β class) and β-lactamase (263 residues, belongs to the class “multi-domain proteins” in SCOP) in four different positions of the host protein phosphoglycerate kinase (415 residues, belongs to α/β class) and showed that both the host as well as the inserted partners are functional.28 They also observed functional coupling between the two fused partners in some of the chimeras. Thus, we believe that our description of the many features of domain insertions could be used for creating novel multi-functional fusion proteins by employing protein engineering methods. We have developed a web resource for domain insertions in protein structures that are classified in the SCOP database†.29

Acknowledgments

We thank Cyrus Chothia and Sarah Teichmann for discussions, Siarhei Maslau and Emma Hill for valuable comments on the manuscript. R.A.-S & R.S are grateful to the Cambridge Commonwealth Trust and the Medical Research Council, UK for financial support.

Abbreviations used

PDB	Protein Data Bank
SCOP	structural classification of proteins

†http://stash.mrc-lmb.cam.ac.uk/DomIns

References

1. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J. Mol. Biol. 1995; 247:536–540. [PubMed: 7723011]
2. Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM. CATH—a hierarchic classification of protein domain structures. Structure. 1997; 5:1093–1108. [PubMed: 9309224]
3. Holm L, Sander C. Mapping the protein universe. Science. 1996; 273:595–603. [PubMed: 8662544]
4. Chothia C. Proteins. One thousand families for the molecular biologist. Nature. 1992; 357:543–544. [PubMed: 1608464]
5. Bork P, Downing AK, Kieffer B, Campbell ID. Structure and distribution of modules in extracellular proteins. Quart. Rev. Biophys. 1996; 29:119–167.
6. Teichmann SA, Park J, Chothia C. Structural assignments to the *Mycoplasma genitalium* proteins show extensive gene duplications and domain rearrangements. Proc. Natl Acad. Sci. USA. 1998; 95:14658–14663. [PubMed: 9843945]
7. Wetlaufer DB. Nucleation, rapid folding, and globular intrachain regions in proteins. Proc. Natl Acad. Sci. USA. 1973; 70:697–701. [PubMed: 4351801]
8. Russell RB. Domain insertion. Protein Eng. 1994; 7:1407–1410. [PubMed: 7716150]
9. Berman HM, Battistuz T, Bhat TN, Bluhm WF, Bourne PE, Burkhardt K, et al. The Protein Data Bank. Acta Crystallog. sect. D: Biol. Crystallog. 2002; 58:899–907.
10. Hobohm U, Sander C. Enlarged representative set of protein structures. Protein Sci. 1994; 3:522–524. [PubMed: 8019422]
11. Chandonia JM, Walker NS, Lo Conte L, Koehl P, Levitt M, Brenner SE. ASTRAL compendium enhancements. Nucl. Acids Res. 2002; 30:260–263. [PubMed: 11752310]
12. Palm GJ, Billy E, Filipowicz W, Wlodawer A. Crystal structure of RNA 3′-terminal phosphate cyclase, a ubiquitous enzyme with unusual topology. Struct. Fold. Des. 2000; 8:13–23.
13. Ditzel L, Lowe J, Stock D, Stetter KO, Huber H, Huber R, Steinbacher S. Crystal structure of the thermosome, the archaeal chaperonin and homolog of CCT. Cell. 1998; 93:125–138. [PubMed: 9546398]
14. Zhang R, Evans G, Rotella FJ, Westbrook EM, Beno D, Huberman E, et al. Characteristics and crystal structure of bacterial inosine-5′-monophosphate dehydrogenase. Biochemistry. 1999; 38:4691–4700. [PubMed: 10200156]
15. Poland BW, Xu MQ, Quiocho FA. Structural insights into the protein splicing mechanism of PI-SceI. J. Biol. Chem. 2000; 275:16408–16413. [PubMed: 10828056]
16. Ichiyanagi K, Ishino Y, Ariyoshi M, Komori K, Morikawa K. Crystal structure of an archaeal intein-encoded homing endonuclease PI-PfuI. J. Mol. Biol. 2000; 300:889–901. [PubMed: 10891276]
17. Jurica MS, Stoddard BL. Homing endonucleases: structure, function and evolution. Cell Mol. Life Sci. 1999; 55:1304–1326. [PubMed: 10487208]
18. Gerstein M. A structural census of genomes: comparing bacterial, eukaryotic, and archaeal genomes in terms of protein structure. J. Mol. Biol. 1997; 274:562–576. [PubMed: 9417935]
19. Doi N, Itaya M, Yomo T, Tokura S, Yanagawa H. Insertion of foreign random sequences of 120 amino acid residues into an active enzyme. FEBS Letters. 1997; 402:177–180. [PubMed: 9037190]
20. Hegyi H, Gerstein M. The relationship between protein structure and function: a comprehensive survey with application to the yeast genome. J. Mol. Biol. 1999; 288:147–164. [PubMed: 10329133]
21. Hua S, Guo T, Gough J, Sun Z. Proteins with class alpha/beta fold have high-level participation in fusion events. J. Mol. Biol. 2002; 320:713–719. [PubMed: 12095249]
22. Mittl PR, Berry A, Scrutton NS, Perham RN, Schulz GE. Anatomy of an engineered NAD-binding site. Protein Sci. 1994; 3:1504–1514. [PubMed: 7833810]
23. Apic G, Gough J, Teichmann SA. Domain combinations in archaeal, eubacterial and eukaryotic proteomes. J. Mol. Biol. 2001; 310:311–325. [PubMed: 11428892]

24. Orengo CA, Bray JE, Buchan DW, Harrison A, Lee D, Pearl FM, et al. The CATH protein family database: a resource for structural and functional annotation of genomes. Proteomics. 2002; 2:11–21. [PubMed: 11788987]

25. Betton JM, Jacob JP, Hofnung M, Broome-Smith JK. Creating a bifunctional protein by insertion of beta-lactamase into the maltodextrin-binding protein. Nature Biotechnol. 1997; 15:1276–1279. [PubMed: 9359111]

26. Duplay P, Szmelcman S, Bedouelle H, Hofnung M. Silent and functional changes in the periplasmic maltose-binding protein of *Escherichia coli* K12. I. Transport of maltose. J. Mol. Biol. 1987; 194:663–673. [PubMed: 2821264]

27. Ay J, Gotz F, Borriss R, Heinemann U. Structure and function of the *Bacillus* hybrid enzyme GluXyn-1: native-like jellyroll fold preserved after insertion of autonomous globular domain. Proc. Natl Acad. Sci. USA. 1998; 95:6613–6618. [PubMed: 9618460]

28. Collinet B, Herve M, Pecorari F, Minard P, Eder O, Desmadril M. Functionally accepted insertions of proteins within protein domains. J. Biol. Chem. 2000; 275:17428–17433. [PubMed: 10747943]

29. Selvam RA, Sasidharan R. DomIns: a web resource for domain insertions in known protein structures. Nucl. Acids Res. 2004; 32:D193–D195. [PubMed: 14681392]

30. Kraulis PJ. MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures. J. Appl. Crystallog. 1991; 24:946–950.

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

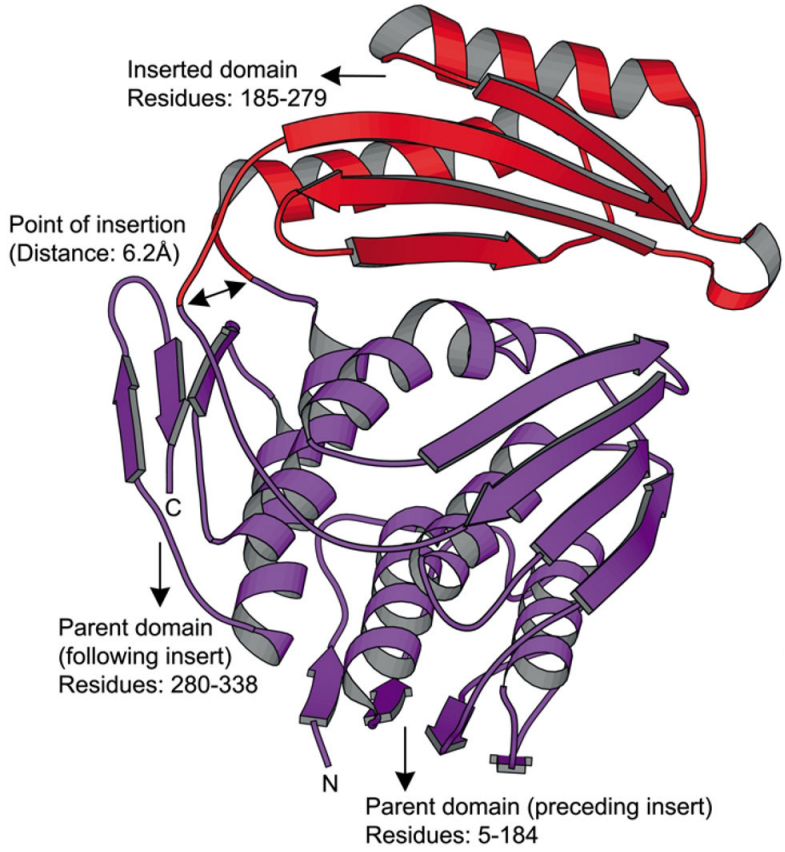

Figure 1.

Domain insertion in *E. coli* enzyme RNA 3′-terminal phosphate cyclase (PDB 1qmhA). The *E. coli* enzyme RNA 3′-terminal phosphate cyclase consists of two domains, of which one is inserted within the other.12 The parent domain (residues 5–184, 280–338, coloured purple) consists of three repeated folding units; each unit has two α-helices and a four-stranded β-sheet. The folding unit resembles the C-terminal domain of bacterial translation initiation factor 3 (IF3). Between an α-helix and a β-strand of the third IF3-like repeat of the parent domain, there is a smaller inserted domain (residues 185–279, coloured red). Although the inserted domain has the same secondary structural elements as the parent domain, it has a different topology and a different fold. Insert resembles the fold observed in human thioredoxin. The figure was prepared using the program MOLSCRIPT.30

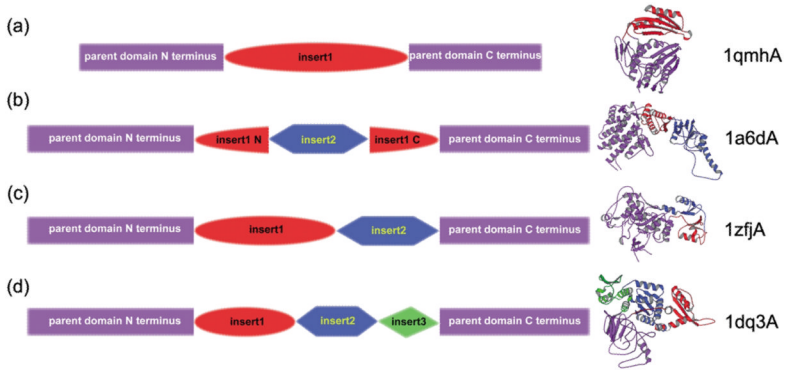

Figure 2.

Schematic representation of types of domain insertions observed in protein structures. Figures of protein structures were prepared using the program MOLSCRIPT.30 (a) Single insertion (e.g., 1qmhA). (b) Nested insertion (e.g., 1a6dA). “insert1 N′ and “insert1 C′ represent the N and C terminus of insert, respectively. (c) Two-domain insertion (e.g., 1zfjA). (d) Three-domain insertion (e.g., 1dq3A).

Europe PMC Funders Author Manuscripts

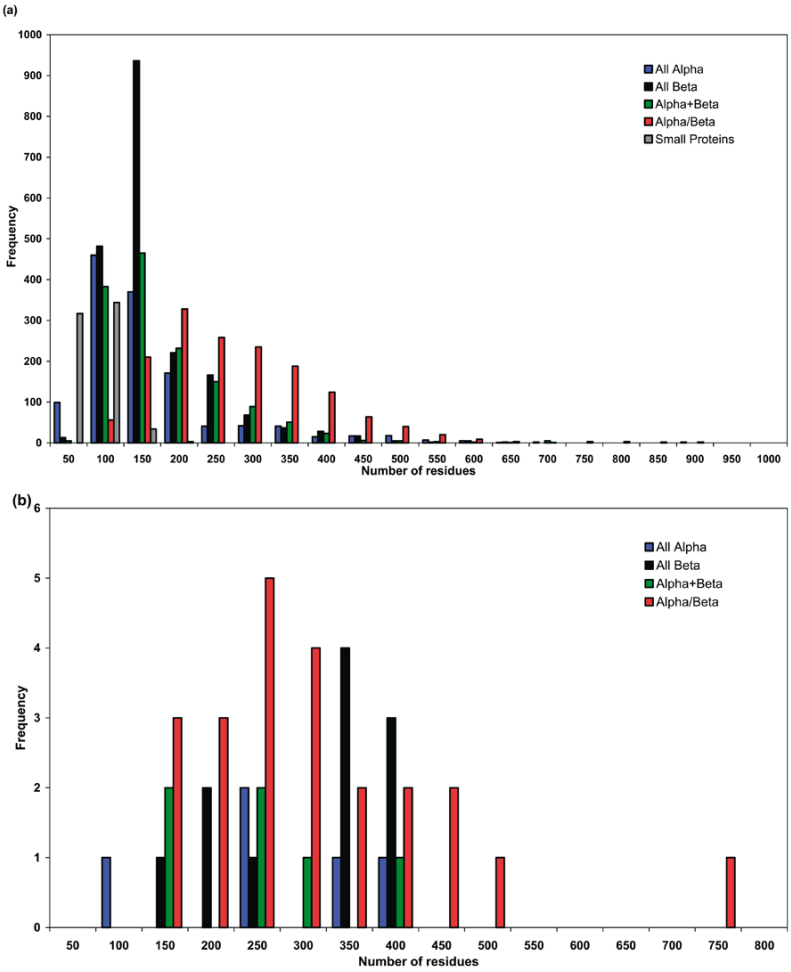

Figure 3.
(a) Domain length distribution for all domains in the non-redundant set of protein structures (PDB_90). (b) Domain length distribution for parent domains.

Europe PMC Funders Author Manuscripts

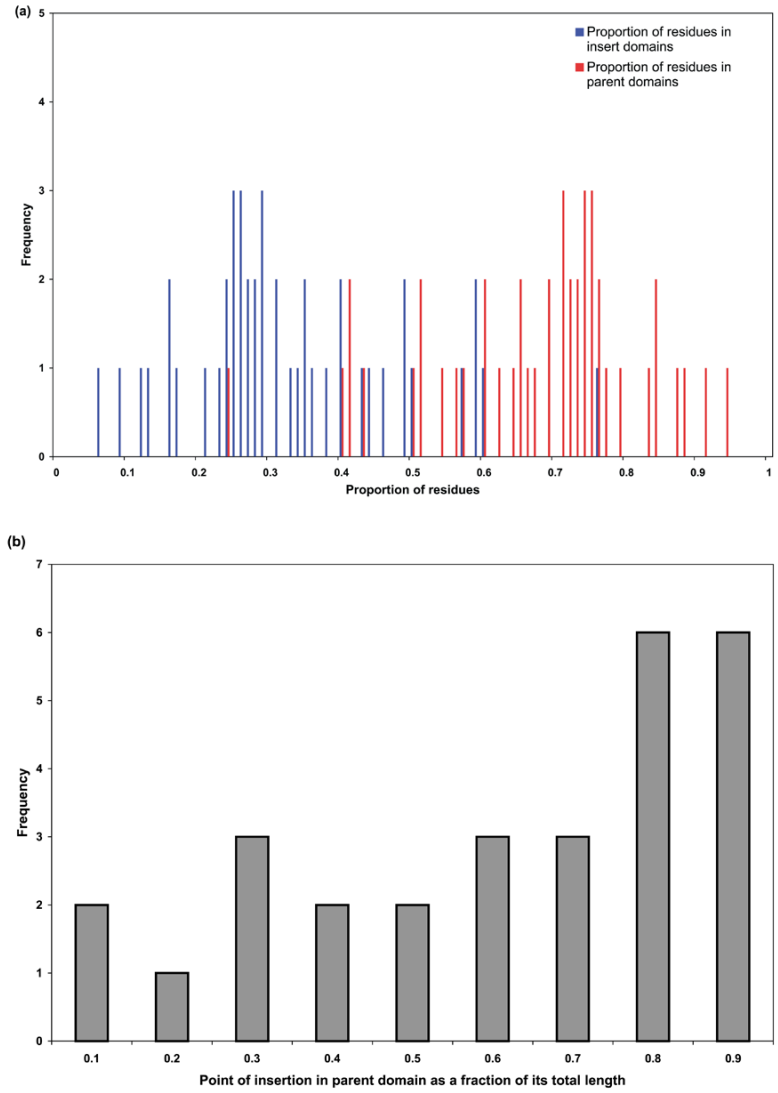

Figure 4.

(a) Proportion of residues in parent and insert domains in parent-insert combinations. (b) Point of insertion in parent domain. Insert position is given as a fraction of total length of parent domain.