APPENDIX A: DOMAIN INSERTION

A.1 Introduction

Taking advantage of an evolutionary basis of domain classification, here I describe the nature and characteristics of domain insertions in protein structures, a phenomenon that is different from the usual pattern of sequential arrangement of domains in multi-domain proteins.

Domains constitute the basic structural, functional and evolutionary unit of proteins (Holm and Sander, 1996; Murzin *et al.*, 1995; Orengo *et al.*, 1997). Proteins can comprise a single domain or a combination of domains. It is well established that multi-domain proteins with widely diversified architecture and functions are generated from a limited repertoire of domain families (Bork *et al.*, 1996; Chothia, 1992). Structural assignments to complete genomes revealed that almost two-thirds of prokaryotic proteins and 80% of eukaryotic proteins are multi-domain proteins (Teichmann *et al.*, 1998). In 1973, Donald Wetlaufer introduced the classification of domains into continuous and discontinuous (Wetlaufer, 1973). 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. Thus, discontinuous domains are essentially formed by one-dimensionally non-contiguous segments of a polypeptide. While most multi-domain proteins have continuous domains, some proteins exhibit non-contiguous arrangement of their domains (Wetlaufer, 1973). In this work, I focus on insertions (Russell, 1994), which are the cases of one domain being inserted into another domain (Figure 67).

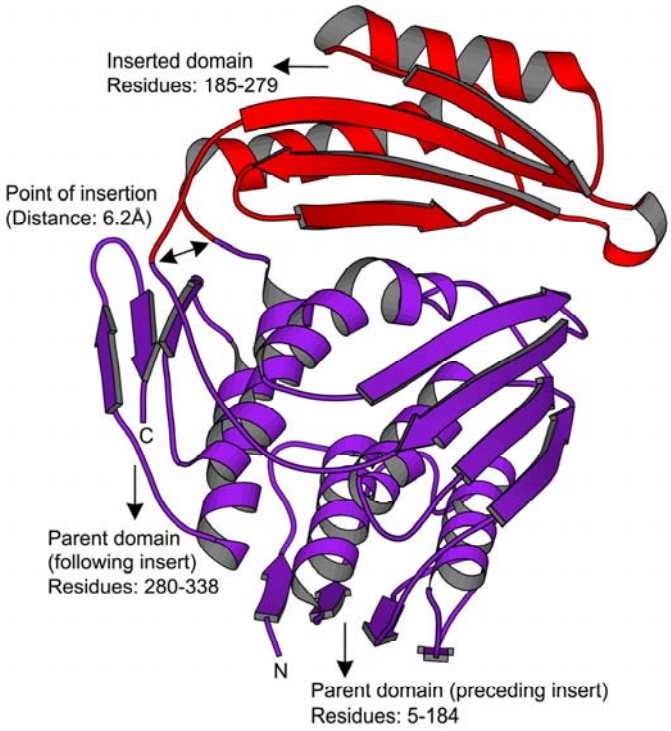

Figure 67. Domain insertion in Escherichia coli enzyme RNA 3'-terminal phosphate cyclase (PDB 1qmhA). The E. coli *enzyme RNA 3'-terminal phosphate cyclase consists of two domains, of which one is contained within the other. The parent domain (residues 5-184, 280-338, coloured purple) consists of three 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 different topology and a different fold. Insert resembles the fold observed in human thioredoxin.*

I followed the definition of protein domains in the Structural Classification Of Proteins (SCOP) database (version 1.61) (Murzin *et al.*, 1995). Although there are several available schemes of protein structure classification, I chose SCOP because it is a manually curated classification of protein structures based on their structural and evolutionary relationship. In SCOP, a protein domain is considered as a unit of evolution if it occurs independently 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 evolutionarily related and can be detected at the sequence level. Domains grouped into 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 my analysis, I considered the fold and superfamily levels of 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 beta sheets (β-α-β units), whereas the α+β class contains proteins with predominantly anti-parallel beta sheets (segregated α and β regions). Small proteins are distinguished by their size rather than other features.

Data for this analysis was obtained from the Protein Data Bank (PDB) (Berman *et al.*, 2002). To overcome the redundancy inherent in PDB, I chose a pre-computed list of non-redundant protein chains provided by PDB_Select (April 2002 release obtained from *ftp://ftp.embl-heidelberg.de/pub/databases/protein_extras/pdb_select*) (Hobohm and Sander, 1994). I used the set of proteins that had pair-wise sequence identities less than 90% and designated this set as PDB_90. Out of the 6182 chains in PDB_90, only 5883 chains were assigned SCOP domain definitions, extracted from the SCOP parseable file *dir.cla.scop.txt_1.61*. Table 24 shows the distribution of SCOP folds, superfamilies, families and domains in each class for chains present in PDB_90.

Table 24. SCOP (1.61 release) classification statistics for chains in PDB_90 (April 2002 release)

Class	Number of Folds	Number of superfamilies	Number of families	Number of proteins	Number of species	Number of domains
All alpha Proteins	147	244	379	719	996	1291
All beta Proteins	109	200	328	784	1475	1981
Alpha and Beta Proteins (a/b)	112	183	434	917	1365	1545
Alpha and Beta Proteins (a+b)	204	287	442	864	1194	1419
Multi-domain proteins	32	32	44	77	124	127
Membrane and cell surface proteins	10	16	28	42	58	120
Small proteins	57	82	123	324	393	698
Coiled coil proteins	4	33	33	48	57	150
Low resolution protein structures	4	4	4	6	6	9
Peptides	40	41	41	59	70	103
Designed proteins	14	14	14	18	18	27
Total	733	1136	1870	3858	5756	7470

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 combinations. I obtained a total of

140 proteins that conformed to this definition. When I considered the 140 pairs of parent-insert superfamily combinations, I observed that several pairs were identical. Whenever there was also the same topological relationship between the parent and insert domains, I retained only one example of a pair of superfamily combinations. This procedure left 40 unique parent-insert superfamily combinations. Variations on the simple scheme 'one insert within one parent' were present; they are shown in Figure 68.

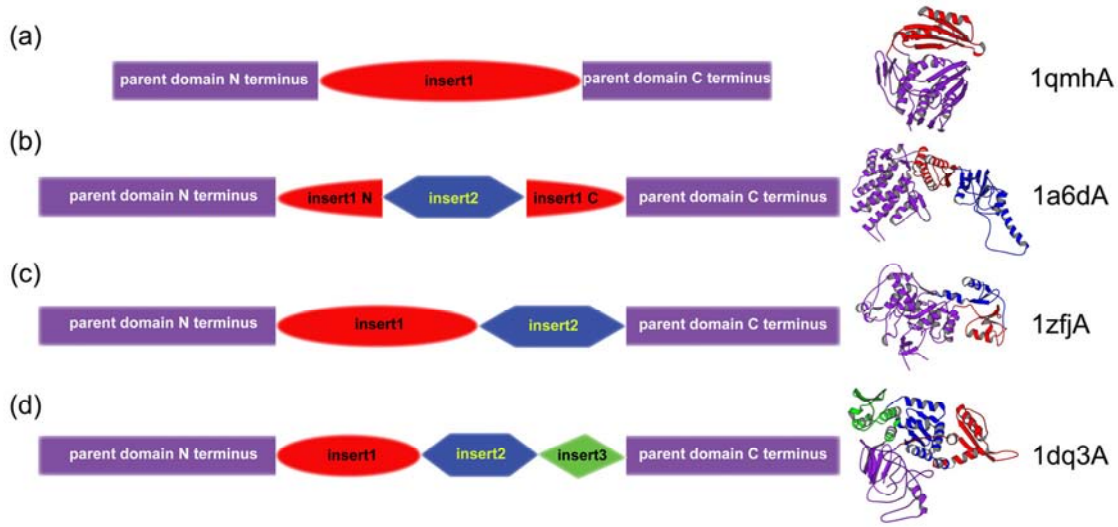

Figure 68. Schematic representation of types of domain insertions observed in protein structures. (a) Single insertion (e.g., 1qmhA). (b) Nested insertion (e.g., 1a6dA). 'insert1 N' and 'insert1 C' represent the N- and C-terminus of insert, respectively. (c) Two-domain insertion (e.g., 1zfjA). (d) Three-domain insertion (e.g., 1dq3A).

For all cases of identified domain insertions, I checked for artefacts arising from missing coordinates. This was necessary because SCOP domain definitions are based on atomic coordinates provided in PDB. To ascertain consistency, I compared atomic coordinates (ATOM records) *versus* sequences (SEQRES records) that were obtained from the ASTRAL compendium (Chandonia *et al.*, 2002). In the majority of cases, sequences were completely covered by coordinates, but in other cases, there were parts of sequences with missing coordinates. However, in none of the latter cases did the absent coordinates obscure the position of inserts.

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

A.2 Types of domain insertions

Domain insertions can be categorized as either single or multiple depending on the number of inserts (Figure 68). 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 68a, the *Escherichia coli* enzyme RNA 3'-terminal phosphate cyclase (PDB: 1qmhA, Palm *et al.*, 2000) has two domains, a small insert and a larger parent that belong to different superfamilies. Close to 90% (36/40) of observed insertions are single insertions. In multiple insertions, more than one domain, either of the same or different superfamily, is inserted into the parent domain. I observed three types of multiple insertions (i) Nested insertions: In *Thermoplasma acidophilum* thermosome (PDB: 1a6dA, Ditzel *et al.*, 1998), the archael chaperonin, the apical domain 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, Zhang *et al.*, 1999) contains two tandem cystathionine-β-synthase domains inserted into the catalytic TIM-barrel domain. The second example is the *Saccharomyces cerevisiae* PI-*Sce*I intein (PDB: 1ef0A, Poland *et al.*, 2000), 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, Ichiyanagi *et al.*, 2000), 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 if a monomer has both copies of the motif (Jurica and Stoddard, 1999). I found that in PI-*Sce*I (case [ii] above) and PI-*Pfu*I (case [iii] above), two monomeric domains were tandemly inserted into one parent domain. The previous observation that motifs are only functional as a dimer suggests 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 this analysis, I treated multiple insertions as several separate parent-insert combinations, resulting in the total of 45 such combinations within 40 protein chains. There were 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, I identified several biologically meaningful patterns. These findings are discussed below.

A.3 Nature and characteristics of domain insertions: Class level

As mentioned before, I considered five SCOP classes, leading to a maximum of 25 (5*5) pair-wise combinations. From the data, I observed only 15 combinations when investigating class participation of parent-insert pairs. The combination of α/β-parent-α+β-insert was predominant, while 50% of all parents belonged to α/β class and 40% of all inserts belonged to α+β class. Domains from α/β class were parent domains, which were two and four fold more often than domains from all-β and all-α class respectively. Domains from the class of small proteins were 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 discussed in the next section.

A.3.1 Size and function of domains involved in insertions

Figure 69a shows the domain length distribution for proteins from PDB_90 set across the five SCOP classes. The average domain length was longest for α/β class followed by the all-β, α+β, and all-α class. When I calculated distribution of average domain lengths for 41 parent domains, I observed the same trend (Figure 69b). However, the average length of parent domains was noticeably larger than the average length of domains from PDB_90 set; this was true for each SCOP class (compare Figure 69a and Figure 69b). 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, I arrived at the explanation that longer domains more readily accept insertions during evolution. As for the inserted domains, α+β and all-α class were equal and major contributors to the number of domains. Therefore, the trend observed for parents is not applicable for inserts.

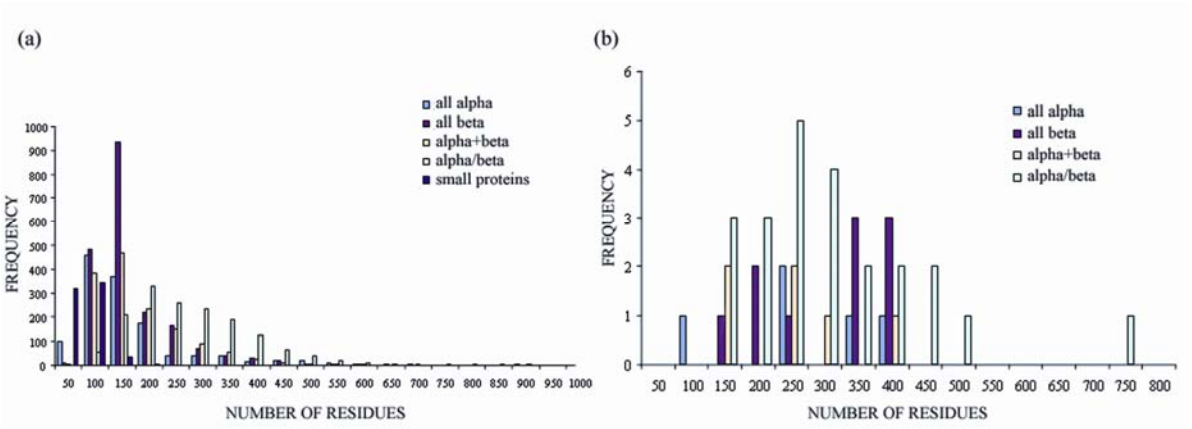

Figure 69. (a) Domain length distribution for all domains in the non-redundant set of proteins (PDB_90). (b) Domain length distribution for parent domains.

In most cases, inserted domains were shorter than parent domains. This is despite the fact that inserted domains could belong to SCOP classes with the longest average domain length (Figure 70a). Parents comprised 50-80% of protein length, while inserts comprised 20-50%. Close to 80% of inserts were shorter than 175 residues, which is the average length of a protein domain calculated from crystal structures (Gerstein, 1997). More than 60% of inserts were shorter than 130 residues. This observation is consistent with the heuristic logic that smaller domains are less likely to disturb the structure and folding of parent domains; it could explain short lengths of inserted domains. This explanation does not contradict an important experiment by Doi and colleagues (Doi *et al.,* 1997). They were able to show that when random sequences of 120-130 amino acid residues were inserted into a surface loop region of *Escherichia coli* RNase HI, about 10% of the clones retained >1% of the wild-type RNase HI activity (Doi *et al.,* 1997).

The high proportion of α/β class domains, as parents, can be correlated with their biochemical function. Previous work showed that more than a half of PDB families 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 (Hegyi and Gerstein, 1999). It is therefore possible to predict that domain insertions are likely to occur in enzymes. Indeed, in the 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 (Hua *et al.*, 2002). α/β folds are also known to be disproportionately associated with enzymatic function (Hegyi and Gerstein, 1999), which lends further credence to the prominent role of α/β folds in accepting insertions.

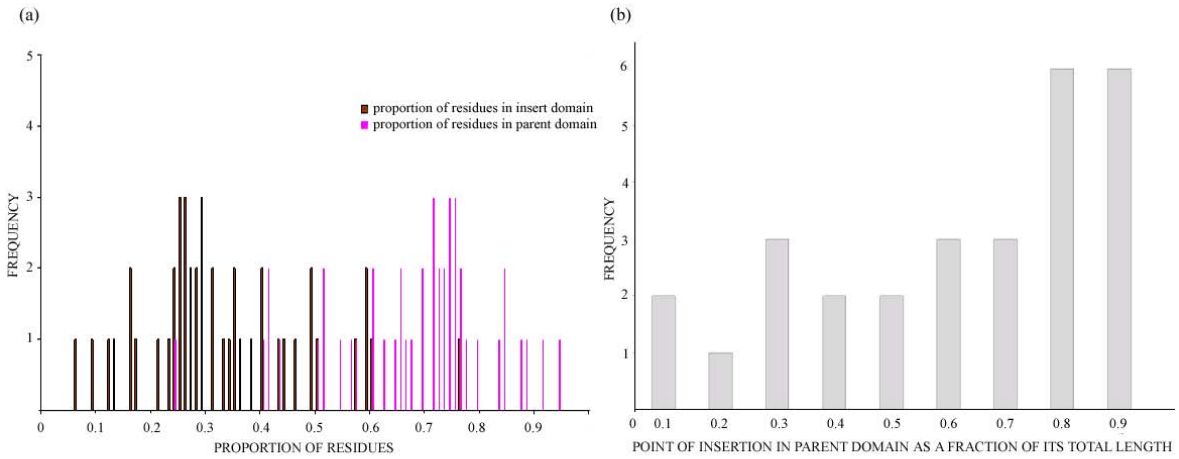

Figure 70. (a) Proportion of residues in parent and insert domains in parent-insert combinations. (b) Point of insertion in parent domain. Insert position is given as a fraction of total length of parent domain.

A.4 Nature and characteristics of domain insertions: Fold and superfamily level

Out of 57 folds in the class of small proteins, two domains with one fold (Rubredoxin fold) were found as inserts; both 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 (Table 25). The trend was the same for the other SCOP classes, where folds of inserted domains constituted minor fractions of all known folds. In contrast to the inserts, all parent domains had different folds. Thus, I observed another distinction between parents and inserts at the fold level.

Similarly, parent superfamilies were found to be more versatile than insert superfamilies (most insert superfamilies combine with only one parent superfamily). There are merely 3 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.

Table 25. Distribution of inserted and parent domains at the SCOP class and fold level. The number of domains and the number of folds they come from is given for inserted and parent domains across the five different classes in the SCOP hierarchy. Percentage gives the number of folds contributing to insertions over total number of folds under the class.

SCOP Class	Total number of folds	Inserted domains			Parent domains		
		Number of domains	Number of folds	Percentage of folds	Number of domains	Number of folds	Percentage of folds
All-α	147	6	5	3.4	5	5	3.4
All-β	109	9	9	8.3	11	11	9.2
α/β	112	10	6	5.4	23	23	20.6
α+β	204	18	11	5.4	6	6	3
Small proteins	57	2	1	1.8	0	0	0

While many parent superfamilies conservatively combine with one insert superfamily, there are conspicuous exceptions. There are three parent superfamilies each combining two different insert superfamilies. The three parent superfamilies in question are Zn-dependent exopepetidases 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 hydrolases superfamily, and FAD/NAD(P)-binding domain superfamily.

Two further observations at the superfamily level are worth mentioning. Firstly, all parents and inserts belong to different superfamilies. There is only one exception: in *Escherichia coli* enzyme glutathione reductase (PDB: 1gesB), the parent and insert belong to the same superfamily of FAD/NAD(P)-binding domains. Secondly, superfamilies that are popular in the parent or insert context also appear to be popular in the sequential domain combination context (Apic *et al.*, 2001). They were found combining with more than one superfamily in the 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.

A.5 Point of insertion

I did not find any bias in the distribution of insertion points within 41 unique parent-insert combinations. However, a significant bias in the location of the insertion point was observed when I 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 70b, 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 were observed in loop regions of the 3D structure of parent domains.

Though it may not be feasible to provide a definitive explanation for the observation of bias towards C-terminus for insertion in the parent domain, an event in the N-terminus or the middle of the domain are likely to disrupt the gene structure and pose a problem during transcription or translation.

Also insertions in the C-terminus indicate most of the insertions seen in the database are not *strictly* insertions but normal sequential combinations with the second domain starting before the end of the first domain. This stem from the fact, C-terminus bias in insertion is found only in cases of parent-insert combinations, where either the parent or insert also occur in sequential combinations with other superfamilies. Further research on the domain insertions involving the core structure of the parent and insert domains can throw more light on this view.

A.6 Proximity of N- and C-termini in inserts

I wanted to determine how the insertion context affects the distance between N- and C-terminus of an inserted domain. The distance between termini was defined as the distance between C-alpha atoms of the first and the last residue of the domain. I first calculated distances for domains that do not participate in insertions. In order to do this, I considered 1000 domains, each representative of one SCOP superfamily. I obtained sequences and coordinates for the domains from the ASTRAL compendium (Chandonia *et al.*, 2002). Only 687 domain sequences were completely covered by coordinates. Using AEROSPACI scores (Chandonia *et al.*, 2002), I was able to find 60 substitutes for the 313 representative domains that were not entirely covered by coordinates. Altogether, I obtained complete coordinate

information for 747 domains (687 + 60). Because I confined the analysis to five major SCOP classes, I calculated distances between termini for the 711 domains, which belong to the five classes being investigated. The average distance for representative domains was 25 Å.

Calculation of distances between the termini of inserted domains was less straightforward. Domain boundaries reported in SCOP are human defined. Therefore, I compared SCOP domain boundaries for 41 inserted domains against the domain boundaries reported in CATH database (Orengo *et al.*, 2002). In contrast to SCOP, CATH structural classification of proteins has been produced automatically. However, only 28 out of 41 inserted domains were available in CATH, whereas the other 13 have either differences in domain classification or the corresponding proteins were absent from CATH classification. For 28 inserted domains, boundaries were identical between SCOP and CATH. The average distance between domain termini of inserted domains was 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 me 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 were 30 Å and 5 Å for parent and an insert, respectively. These figures were 11 Å and 8 Å for NAD-binding Rossmann domain superfamily. Thus, this analysis shows that the ends of inserted domains are significantly closer than ends of parent domains or domains not participating in insertions. However one must be cautious in interpreting the results as the N and C termini distances for the parent domain is not calculated for the core structure.

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.

A.7 Conclusions

Utilising an evolutionary basis of domain classification, I described the nature and characteristics of domain insertions in protein structures. Domain insertions represent an unusual but abundant case of multi-domain proteins. This analysis gave several novel insights into the nature and characteristics of domain insertions.

(1) Close to 9% multi-domain proteins contain insertions.
(2) The majority of insertions are the single domain insertions. Also found there were two-domain, three-domain, and nested insertions in PDB.
(3) α/β class has a higher propensity to accept insertions. This could be correlated to the size and function of proteins within the class.
(4) Parent domains were found to be longer than the inserted domains in most cases.
(5) When fold and superfamily combinations were considered for parents and inserts, the former was found to be more versatile than the latter, in that the parent domains combined with more 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 have juxtaposed termini compared to parent domains.

Perhaps, domains are more viable in the insert context when their termini are close in space; small size can further contribute to their viability.

These results clearly indicate that despite the structural and functional constraints inherent in the process of domain insertion, this process is an effective way of creating multi-domain proteins. This description of the many features of domain insertions could be used in protein engineering for producing novel multi-functional fusion proteins. Betton and co-workers (Betton *et al.,* 1997) created hybrid proteins by inserting a penicillin-hydrolysing enzyme TEM beta-lactamase (Bla) into the maltodextrin-binding protein (MalE); they used the permissive insertion sites identified before (Duplay *et al.,* 1987). Two insertions resulted in the 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, and the authors experimentally showed the 5 Å distance between the termini of the inserted

domain (Bla). Thus, there is recent experimental data that nicely fit into the picture of insertions found in natural multi-domain proteins.