APPENDIX B: PROTEIN EVOLUTION

B.1 Introduction

Divergence in structure and function of proteins is due to an evolutionary process driven by functional and environmental constraints. These constraints bring about changes in the protein sequence through mutations, insertions and deletions with the preservation of residues important for the structure and function of the protein (Chothia and Lesk, 1986). However, not all the sequence modifications are incorporated or maintained since some changes may be deleterious to the structure or function of the protein. Hence, the structural *'core'* (Chothia and Lesk, 1986) tends to be well conserved during evolution. When proteins evolve, the constraints on the protein structure are relaxed or rather replaced by new constraints and the sequence and structure can change more radically. These changes are generally slow processes and leave a trail of *homologs*. Homologs are proteins evolved from a common ancestor and their evolutionary relationship is evident from similarities in sequence, structure and function. Homologous proteins have been studied for a long time to understand their evolutionary relationships and to assign function or structure to new protein sequences. For homolog searches in the sequence databases, one needs an alignment algorithm, residue similarity matrix, scoring scheme and knowledge about scoring thresholds to identify true relationships.

Among the available pairwise alignment algorithms, one of the most sensitive is the Smith-Waterman algorithm (Smith and Waterman, 1981) adopted in the SSEARCH program (Pearson, 1991). Although this algorithm is more sensitive and rigorous, it is computationally expensive in comparison to FASTA (Pearson and Lipman, 1988) and BLAST (Altschul *et al.*, 1990). The speed and convenience of BLAST made it the most popular program, although it compromises sensitivity. FASTA ranks between these two programs and can be run in two modes: either at greater speed (ktup = 2) or greater accuracy (ktup =1). Pearson (Pearson, 1991, 1995) did a comparison of these three methods and showed that the Smith-Waterman algorithm worked slightly better than FASTA, which was in turn much more effective than BLAST.

Although pairwise comparison methods are a common way to find sequence homologs, they have difficulty in detecting remote homologs when sequence identity falls below 30% (Brenner *et al.*, 1998). Alternate methods like Profile Hidden Markov Models (Eddy, 1996; Krogh *et al.*, 1994), psi-BLAST (Altschul *et al.*, 1997) and Intermediate Sequence Search (Park *et al.*, 1997) reduce this limitation and increase sensitivity.

Intermediate Sequence Search (ISS) is a search technique, wherein two related sequences which cannot be detected directly by pairwise sequence comparison methods are matched using an intermediate sequence sharing close homology with the two distantly related sequences. This concept has been extended to include multiple intermediate sequences (MISS) between two distant sequences (Salamov *et al.*, 1999). The disadvantage with ISS is that the errors caused in the intermediate are likely to propagate as it is not dependent on multiple sequence alignment. Errors caused by ISS when comparing multi-domain protein sequences, can be avoided by splitting query sequence to individual domains. Figure 71 gives an overall idea on how different methods are exploring the sequence space (Lindahl and Elofsson, 2000).

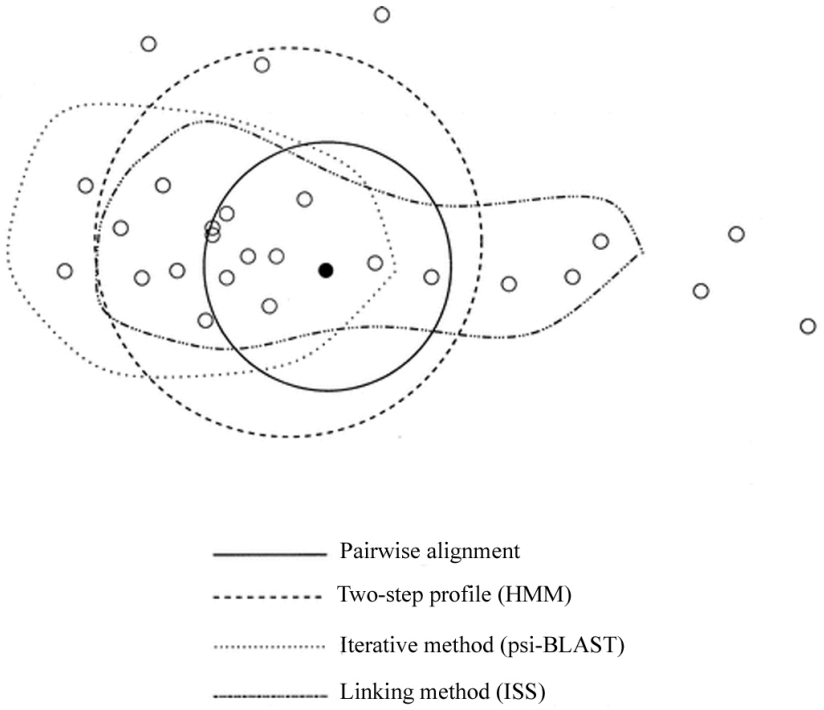

Figure 71. Schematic diagram showing performance of different sequence comparison methods. The filled circle represents the query sequence used in the database search and the open circles represent family members. The distance between two circles represents some arbitrary distance.

A comparison of these recent methods with pairwise sequence comparison methods, performed by searching remote homologs in a Structural Classification Of Proteins (SCOP, Murzin *et al.,* 1995) sequence database having less than 40% identity, show that ISS performs one and half times better than FASTA. In sequences with less than 30% identity, a HMM-based SAM-T98 and psi-BLAST detected three times more relationships than pairwise sequence comparison methods (Park *et al.,* 1998). Sauder *et al.* compared the quality of alignments produced by BLAST, psi-BLAST, ISS and ClustalW (Thompson *et al.,* 1994) with structural alignments. ISS produced longer alignments than psi-BLAST with nearly comparable per-residue alignment quality. At 10-15% identity, BLAST correctly aligned 28%, psi-BLAST 40% and ISS 46% of residues to the structural alignment (Sauder *et al.,* 2000).

All these results show that ISS performs as well as psi-BLAST in identifying distant homologs. However it is not yet clear how ISS is able to detect remote relationships. Moreover, I was interested to determine whether intermediates identified by ISS can provide any knowledge about protein evolution. This study tries to find answers to these questions.

To aid this objective, I also used structure comparisons to understand relationships between proteins. The degree of fitness between structures is usually calculated by a scoring scheme. The common way to represent the structural fitness is Root Mean Square Deviation (RMSD) for all residues of the two protein structures. The RMSD gives a measure of the average level of deviations over the superposed atoms.

$$\sqrt{\sum_{i=1}^{n} \frac{D_i^2}{N}}$$

Where, *D* refers to deviation of the atoms and *N* refers to the number of atoms matched.

There are different structural alignment methods adopting the aforementioned algorithms. Amongst the common implementations are DALI (Holm and Sander, 1993), Combinatorial Extension (CE) (Shindyalov and Bourne, 1998), and Protein Informatics System for Modelling (PrISM) (Yang and Honig, 2000). Here, I used PrISM to compare the structures.

Protein evolution may occur in two ways: divergent or convergent evolution. When a protein structure diverges to form a new fold or function, it results in divergent evolution

(e.g., P-loops). However if two evolutionarily independent folds converge to represent similar structure or function it becomes convergent evolution (e.g., serine proteases). Proteins evolved through a divergent mechanism are likely to have a trail of homologs and can be detected using sequence and structure comparisons. Here, I attempt to study this using two well known protein families – *Cytochrome c* and *P-loops* and answer the following questions.

(1) Is it possible to understand the evolutionary pattern of any protein family or superfamily based solely on its structure and sequence divergence?
(2) Whether understanding this will help us in assigning hierarchies for a protein in the existing classification of protein structures?

B.2 Datasets

I used SCOP database for this study (please refer to Appendix A for details of SCOP). The *All-α* protein class contains a fold level called *cytochrome c*, which in turn is composed of a single superfamily named *cytochrome c*. This superfamily has four families. The *Di-haem cytochrome c peroxidase* family has only synthetic protein structures and, therefore, only domains from the other families (39 sequences) were used in this analysis.

P-loop domains are found in the class *α/β* and fold/superfamily *P-loop containing nucleotide triphosphate hydrolases* (this fold has only one superfamily). The superfamily has domains composed of parallel beta sheets of varied sizes connected by helices. For example, the *Nucleoside and nucleotide kinases* family has 5 strands with architecture type 23145 and *Nitrogenase iron-protein like group* family has 7 strands with architecture type 3241567. The superfamily is composed of 14 families. I used all the domains (85 sequences, excluding domains involving multiple chains) from these 14 families for this analysis.

From these datasets, I then found sequence homologs and structure homologs that can be detected by the above described methods.

B.3 Intermediate sequence search

I collected homologs for each of the domains in the two superfamily datasets using FASTA 3.3 (with BLOSUM 62 matrix, ktup = 1) by searching against the pdb90d_1.53 database. The pdb90d_1.53 database is derived from sequences of SCOP domains (version 1.53) sharing 90% or less sequence identity.

Domains (query and target), with scores better than the threshold value 0.01, are referred as 'direct hits'. For domains that cannot be detected directly, I used the ISS procedure described above to link the query and target.

A comparison of ISS hits with psi-BLAST shows that psi-BLAST can detect all the remote homologs identified by ISS in P-loops superfamily and only about half of them in cytochrome c superfamily. The advantage ISS has in some cases might be due to the match score it gains by producing longer alignments around conserved regions of the protein. However, both the methods fail to detect remote homologs from P-loops superfamily than found from cytochrome c superfamily. This might be due to the extensive divergence of sequences in P-loops superfamily (they are quoted to have some converged domains (Bossemeyer, 1994) and differences in sequence length (average length of P-loops is ≈ 230 amino acids, twice the size of cytochrome c).

Intermediate searches based on structural information could find new remote homologs that ISS could not detect. This is expected because it is known that different sequences can have similar folds. Therefore, by comparing structures it is more likely to detect remote homologs. I suggest that by using intermediate structural search, even more distant relationships can be detected.

Then I used the alignments obtained from the query-intermediate and target-intermediate to generate a "progressive alignment" (i.e., a multiple sequence alignment generated by progressively aligning pairwise alignments using *ClustalW* alignments and structure information) of query-intermediate-target or query-intermediate-intermediate-target.

These progressive alignments show that the intermediates can improve the quality of alignments between query and target. An example of this alignment is shown in Figure 72.

The figure shows the improvement in alignment between query-target (SCOP Ids: *d1a56__* - *d1c75a_*) produced by FASTA (Figure 72a) and the progressive alignment generated manually after introducing one (*d451c__*) and two intermediate (*d1ayg__* and *d451c__*) sequences (Figure 72b and Figure 72c). The alignment shows that there are some residues common in all the sequences and some between query-intermediate, target-intermediate and intermediate-intermediate.

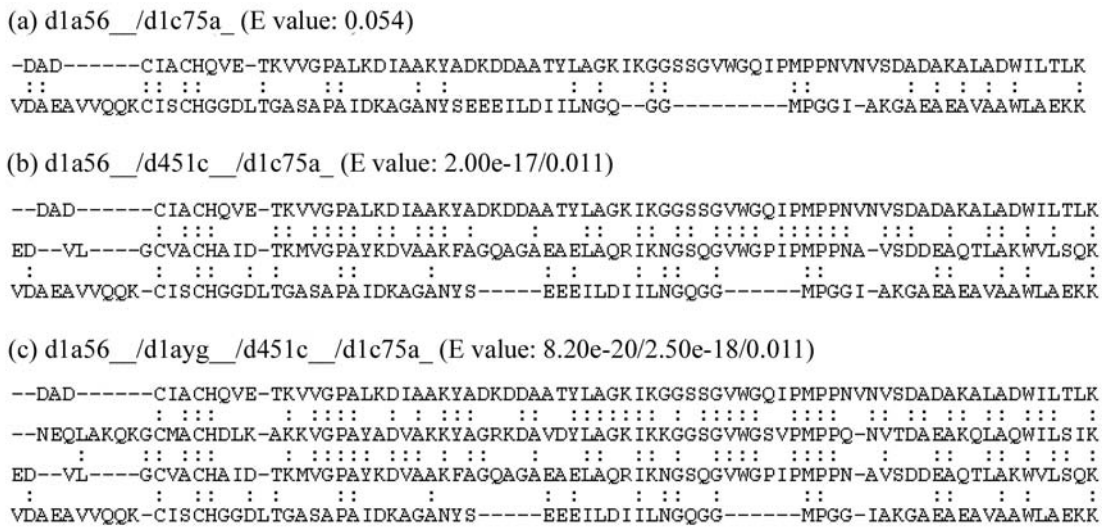

Figure 72. Comparison of alignments of two distant proteins with and without intermediates. (a) Alignment of the two domain produced by FASTA 3.3. (b) The progressive alignment generated by including one intermediate. (c) The progressive alignment generated by including two intermediates.

Likewise, I selected closely clustered domains from each of the four SCOP protein groups (*mitochondrial cytochrome*, *cytochrome* c_2, *cytochrome* c_{551} and *cytochrome* c_6) to make a progressive alignment. These groups were used due to the fact that they represent most of the members of the superfamily. From the progressive alignment made for each of the protein groups, I derived a consensus (Figure 73). This consensus was then used to derive an overall consensus shown in Figure 74. The figure shows that there are 10 invariable residues in the consensus and it agrees with the consensus derived by Ptitsyn by aligning 164 sequences from the cytochrome c superfamily (Ptitsyn, 1998). His alignments were generated using the PileUP program and manually edited taking functional residues into consideration.

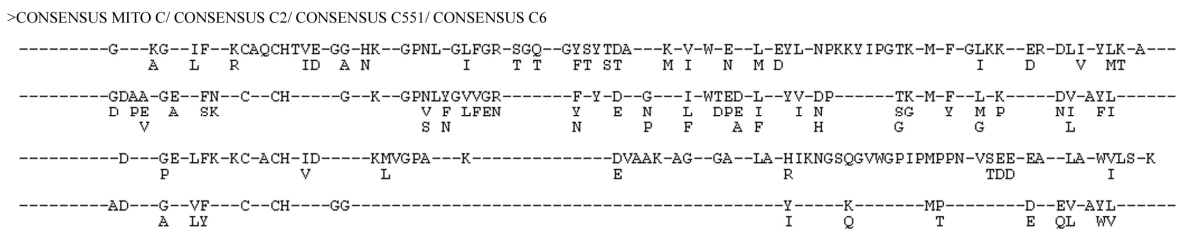

Figure 73. Consensus sequences derived for the four SCOP protein group in monodomain cytochrome c family

Figure 74. Consensus of consensus for sequences in monodomain cytochrome c family

The conserved residues were involved in heme binding and needed for functional role of the protein. The other conserved residues do not have any functional role and are found to be key residues needed to maintain structural fold of cytochromes. The key residues reported here agree well with the results found in the literature (Ptitsyn, 1998). Figure 74 shows the key residues identified by Ptitsyn. The differences include two additional residues conserved at position 3 (aliphatic residue) and position 10 (aliphatic residue), the presence of a proline at position 1 and a phenyalanine instead of an isoleucine at position 8. These discrepancies might be due to number of sequences compared and the kind of alignment generated. Ptitsyn used 164 sequences whereas here only 19 sequences were used. Although comparatively very few sequences were used, the result seems to be almost the same. This is a promising result opening opportunities in extending the procedure to other superfamilies. However, an attempt on P-loops failed primarily due to the fact that the superfamily is much more diverged and only very few sequences form distinct clusters.

B.4 Structural homologs

I did an all-against-all structural comparison of the domains using PrISM. Then I used the alignment from PrISM as input to another program called MSARMS (Hubbard, 1994) that measures the distance in Angstrom between the matched residues in the superposition. These RMSD values from PrISM and MSARMS programs were used for this study.

B.5 Clustering

With these homologs and their relationship (given as *E-value* for sequences and *RMSD* for structures), I represented proteins as clusters in two-dimensional space. This was done using the procedure given in Figure 75 using sequence/structure distance matrices (or similarity matrices).

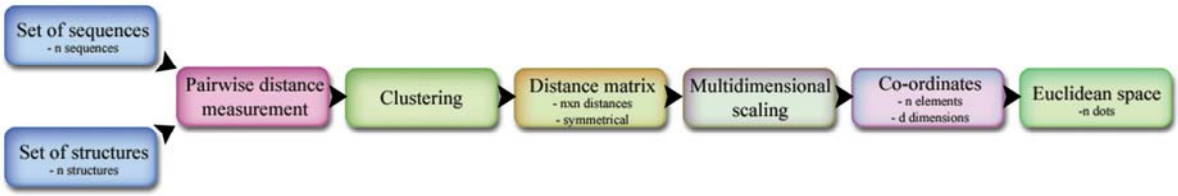

Figure 75. Flow chart describing steps used in clustering and visualisation of data.

I did initial clustering based on the sequence based distance matrix using single and complete linkage methods with a threshold E-value of 0.001 and 0.05 respectively. Then I merged the resulting sets of clusters based on the RMSD values using the Unweighted Pair Group Method using Arithmetic average approach. A threshold value of 4.00Å was used for the P-loops superfamily and a threshold of 2.00Å was used for the cytochrome c superfamily. I also applied the complete linkage approach to merge the initial set of clusters using a threshold value of 6.00Å for both superfamilies.

To find co-ordinates of the data set in 2D space, I used Principal Co-ordinate Analysis (PCoA). For a problem of *N* objects, there could be *N*(N-1)* distances and displayed in *(N-1)* dimensional space. This *(N-1)* dimensional space was reduced to 2D/3D space and plotted.

A manual plotting of the data gave a cluster map for both cytochrome c (Figure 76) and P-loops superfamilies (Figure 77). Figure 78 shows the demarcation of clusters into family and protein levels based on the SCOP classification for cytochrome c. Similarly, Figure 79 shows the demarcation of family levels in P-loops. The protein levels were not marked in P-loops to avoid the complexity in the figure.

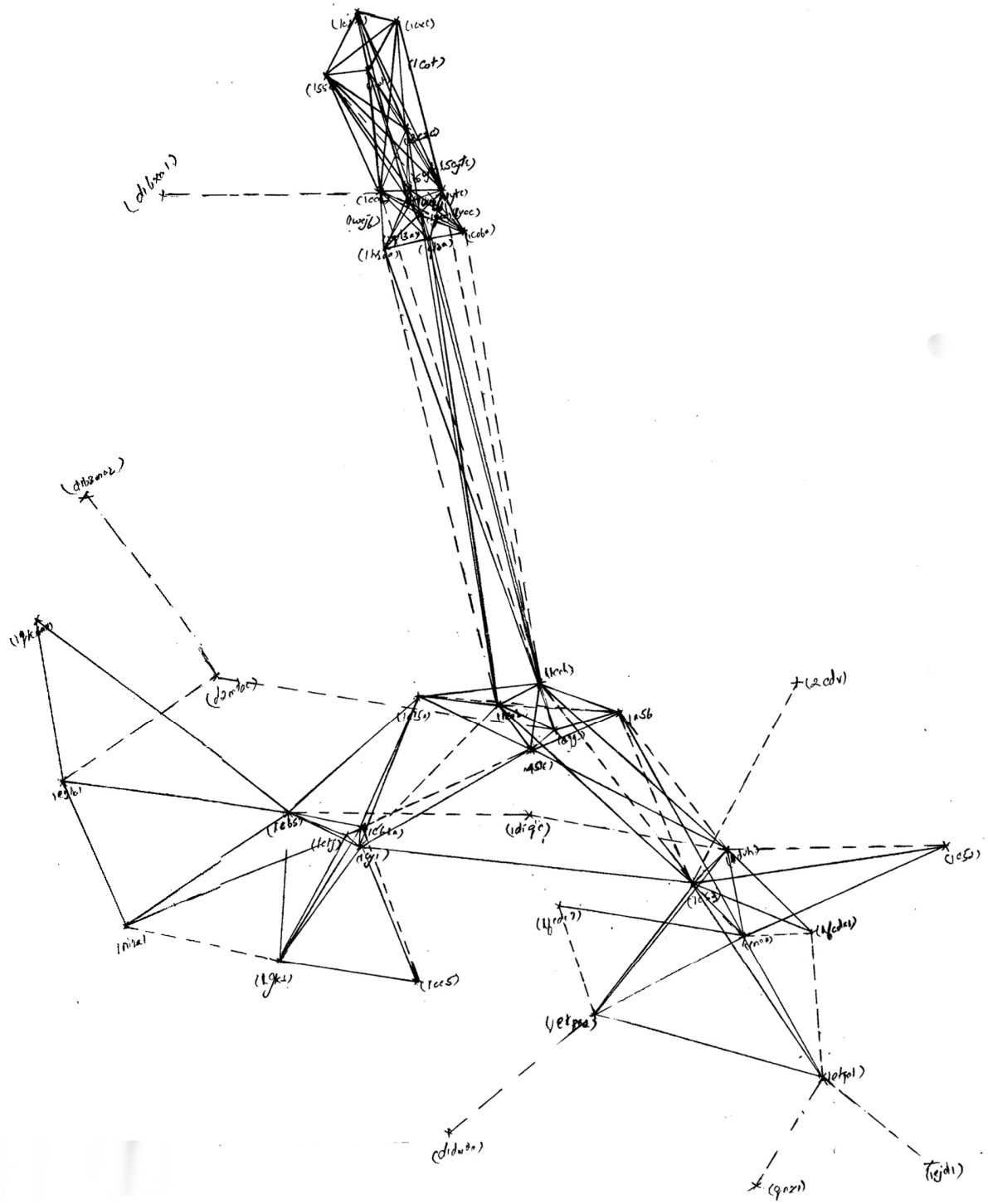

Figure 76. Cluster map of cytochrome c superfamily

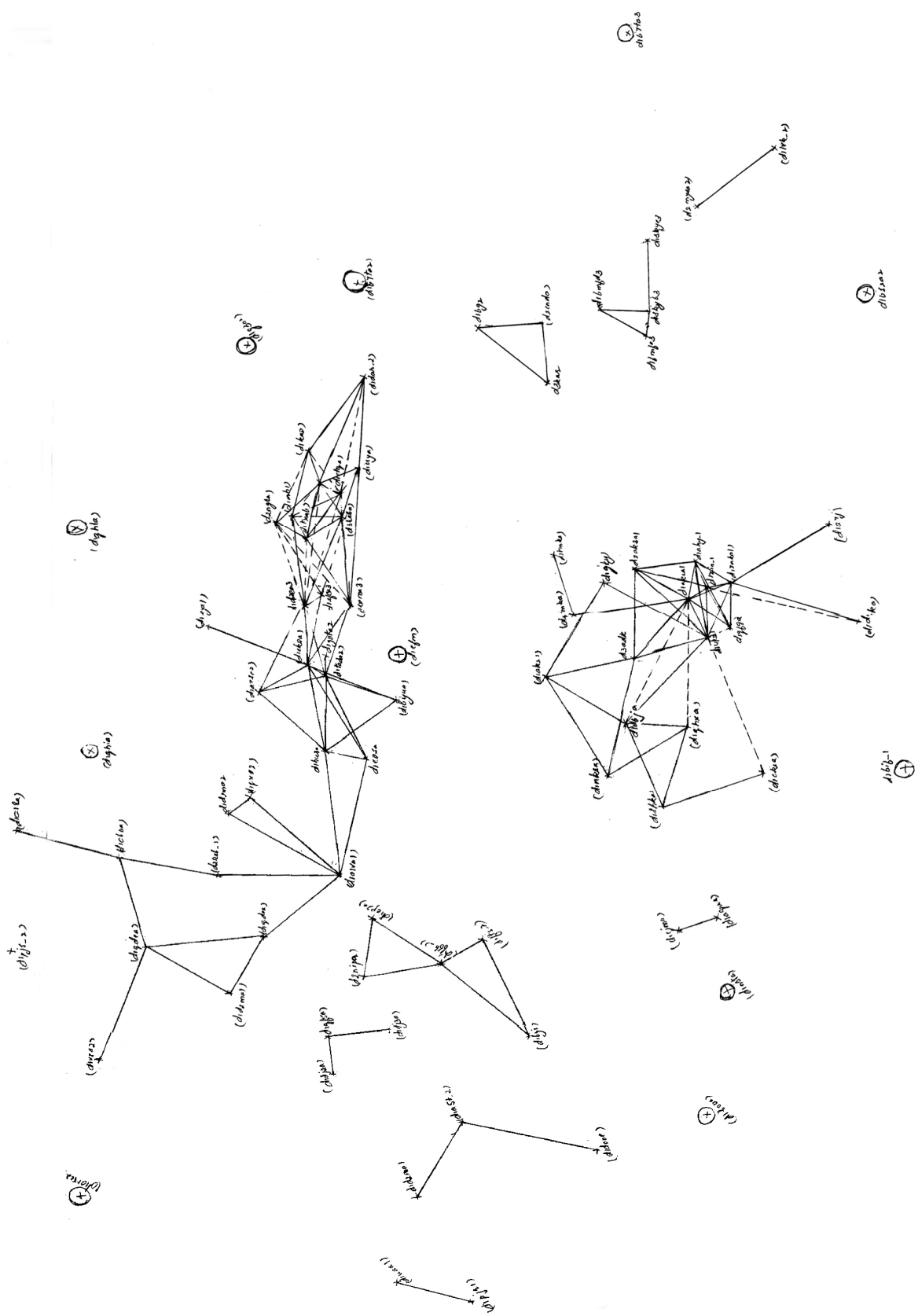

Figure 77. Cluster map of P-loops superfamily

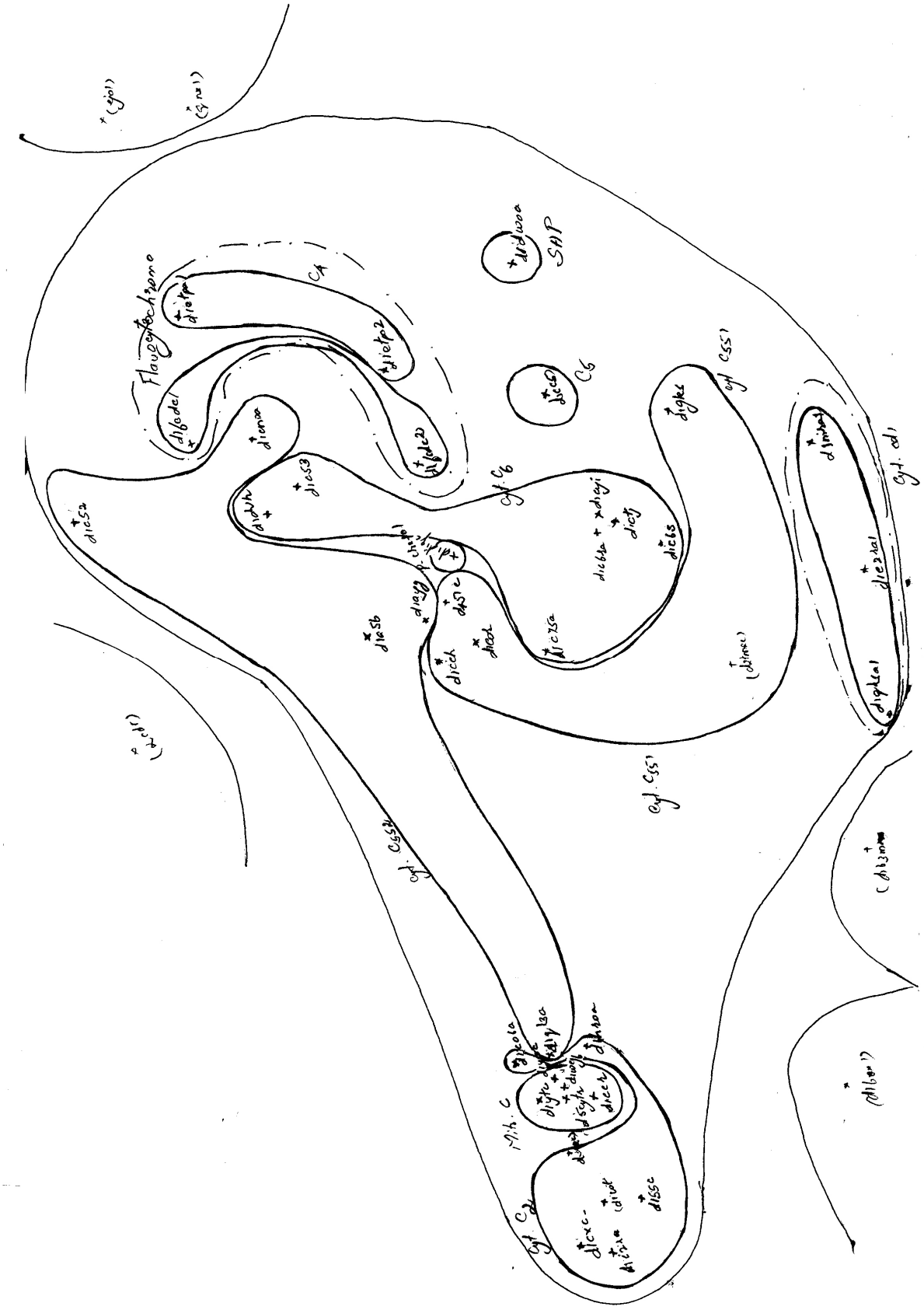

Figure 78. Cluster map of cytochrome c superfamily with demarcation of SCOP superfamily, family and protein levels

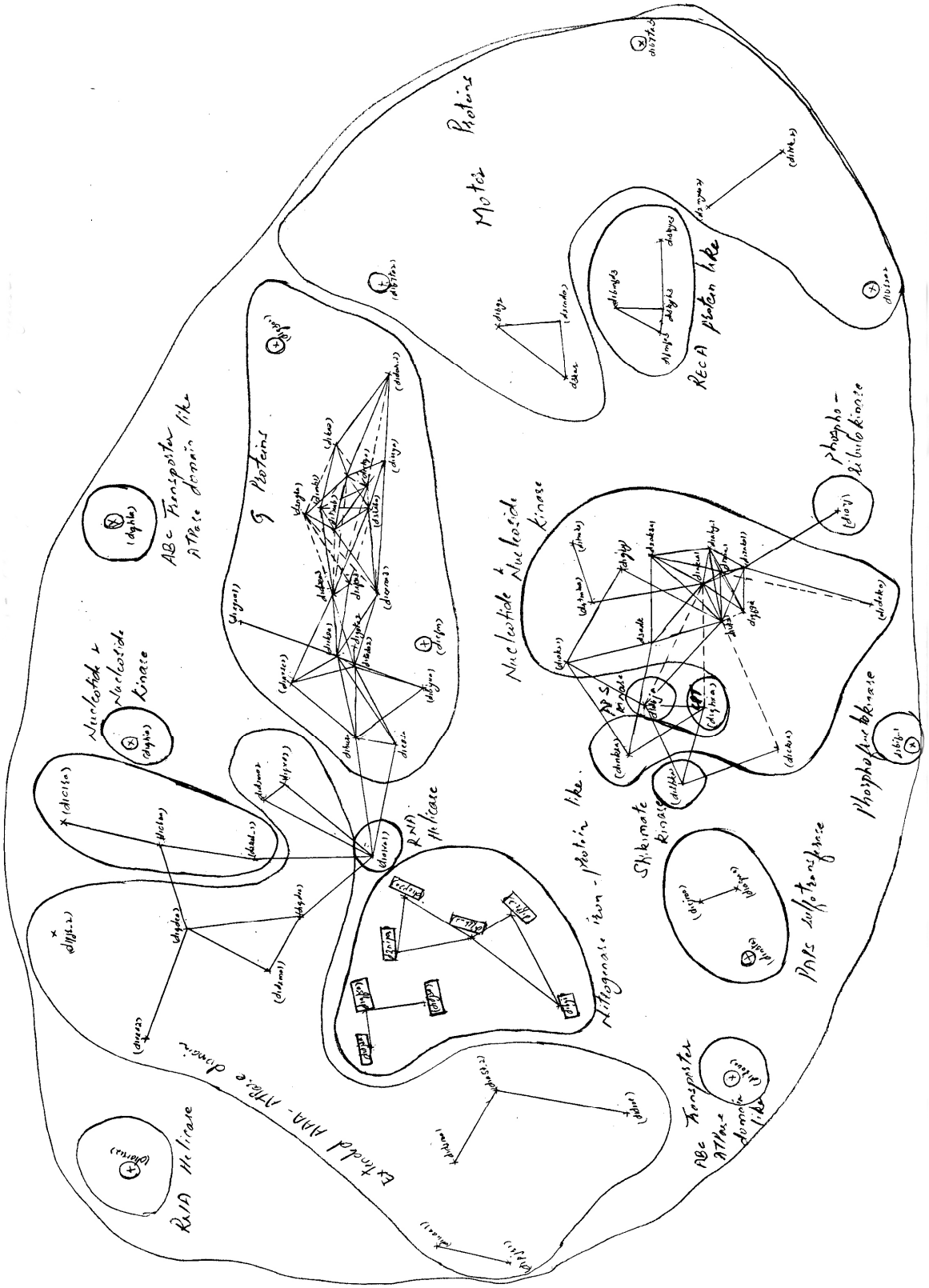

Figure 79. Cluster map of P-loops superfamily with demarcation of SCOP superfamily, family levels

The maps (Figure 76 and Figure 77) show domain relationships either by solid lines or dashed lines. The solid lines indicate domains having strong relationship between them (E-value < 0.4 and RMSD < 4 Å). Also, the length of the solid line represents real Euclidean distance in the cluster map. The dashed lines show there is a relationship between the connected domains. However, the position of domains in the map is not true. This is due to the non-availability of a relationship between the connected domains and its neighbors. Also, the length of the broken line does not represent real Euclidean space in the map.

The cytochrome maps (Figure 76 and Figure 78) show that two SCOP protein groups, *mitochondrial cytochrome c* and *cytochrome* c_2, were well separated from other protein groups. The domains forming the *cytochrome* c_{552} cluster show that they have diverged more than any other SCOP protein group. Also, it can be seen that most of the domains from the *cytochrome* c_6 and *cytochrome* c_{551} SCOP protein groups form closer clusters while some of them get away from this cluster and act as outliers.

P-loops cluster maps (Figure 77 and Figure 79) show that the domains have diverged more when compared to the cytochrome c domains. The maps show a number of domains represented as singletons or as small groups not connected to each other. As stated earlier, absence of a line between domains means no relationship can be identified among them (with score below the threshold limit), although some of the singletons belong to SCOP family. Only members of two families (*Nucleoside and nucleotide kinase* and *G-proteins*) were found to be grouped together on the map. This may be due to more environmental constraints and less active site requirements on P-loop superfamily or may be due to a convergence phenomena as seen in phosphate binding proteins (Bossemeyer, 1994).

These cluster maps are a useful tool to aid in understanding of the relationship between protein members of a family:

(1) It gives an overall picture of the divergence of a protein superfamily.
(2) It shows the relationships between SCOP families.
(3) The method could be used as an initial automated classification procedure of protein structures. A new protein structure can be used as a query to find its sequence or structure homologs. Then based on the sequence and structural relationship (E-value and

RMSD), the protein can be added in the cluster map. Such a map will give a good idea to which of the superfamily or family the new protein belongs. Then with detailed knowledge, the protein can be allocated in a specific family (manual curation). The clustering approach can be exploited to assign function to an unknown protein (Sternberg, 2001), but it cannot be trusted fully as a similar structure does not always represent the same function.

(4) It gives a clear picture about any particular SCOP family and allows the identification of any outliers in it. In the P-loops cluster map (Figure 79), there are two clusters one with domains *d1d2ja__, d1qf5a__ and d1dj3a_* and another with *d2nipa_, d1cp2a_, d1ffh__, d1byi__ and d1fts__* (boxed). But all of these domains are placed in the same family in SCOP. On discussion with Alexey Murzin (the primary curator of SCOP database), he recalled he considered that it might be better to keep these two clusters in two separate groups, say as, two different sub-families/families. He only kept them together due to limitations in the current SCOP classification system.

Likewise the domain *d1qhia_,* classified in the *Nucleotide and nucleoside kinase* family in SCOP, are positioned separately from the main cluster. The outlier was later cross-checked with structural analysis (Morea, 2001). The analysis also agreed that the domain is distinct from its family members. The probable reason for the isolated cluster of *d1qhia__* is that it is a chimeric protein and does not exist naturally i.e. it does not have sequence or structure homology with other *Nucleotide and nucleoside kinase* proteins even though it retains the same function. It was for this reason and since the domain satisfied minimal the P-loop topology, that Alexey Murzin classified the domain under the same family.

Thus, cluster maps might help us to be aware of outliers in a particular superfamily/family classification before starting any kind of detailed analysis on it.

Because of these advantages of the cluster maps, I automated the clustering process to extend the study later for other families. A comparison between manual and automated clustering procedures shows that the automated method performed equally well with the manual method (Figure 80 and Figure 81). Also, the automated methods provide similar results with another automated clustering procedure based on the MCL algorithm (Enright *et al.*, 2002).

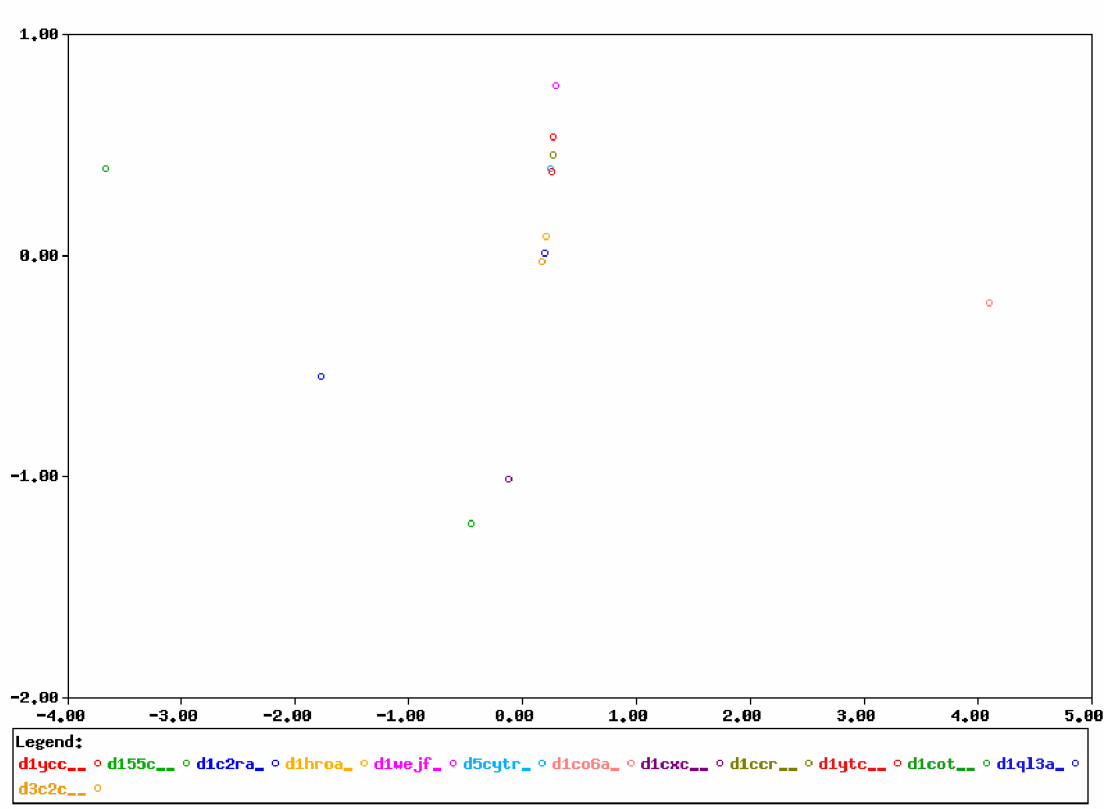

Figure 80. A cluster produced by the automated method for cytochrome c superfamily

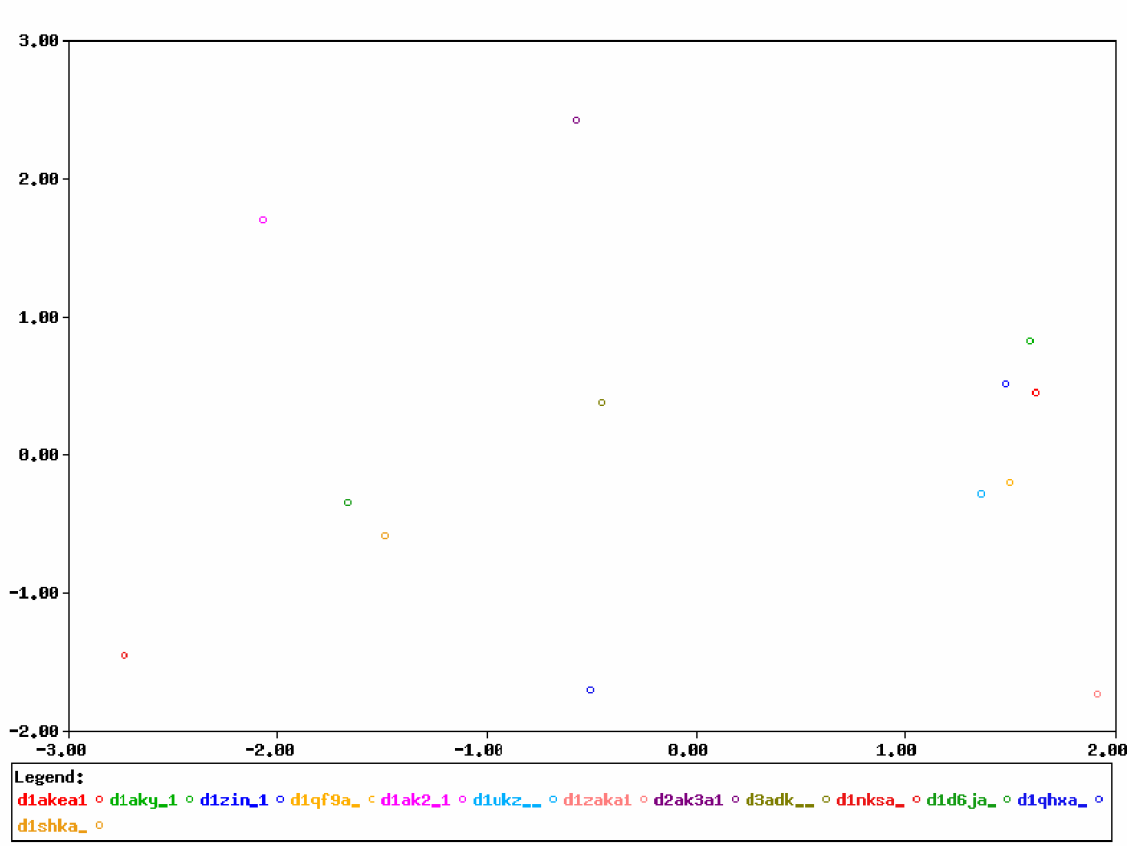

Figure 81. A cluster produced by automated method for P-loops superfamily

In both manual and automated processes, clustering was done using sequence and structural relationships, but it is possible to be done with sequence information alone. However, this will give only the number of clusters that can be formed from the superfamily and members in each cluster. A two dimensional representation of data is difficult with sequence information alone due to the fact that the data needs to undergo significant normalization procedures before it can be used to find co-ordinates.

B.6 Orthology and paralogy

The sequence and structural information, used above to generate cluster maps, can also form the basis for detecting orthologous relationships within protein families in the study of protein evolution. Such a group of ortholog domains was found in P-loops superfamily. The group comprises adenylate kinases from *Escherichia coli*, *Bacillus stermathermophilus* and *Saccharomyces cerevisiae*. Using species as a time scale, it can be said that adenylate kinase of *Escherichia coli* and *Bacillus stermathermophilus* appeared earlier than yeast protein. However, it does not mean that yeast protein evolved from *Escherichia coli* or *Bacillus* and it would be extremely difficult in assessing the proper time scale for these proteins based on sequence and structure information alone.

All the three adenylate kinases clustered close to each other on the map. So, from tightly clustering domains, it can be presumed that they are possibly to be orthologous to each other.

The TOPS (Westhead *et al.,* 1999) diagrams of these three proteins (Figure 82) shows that *Escherichia coli* and yeast adenylate kinases are identical whereas in *Bacillus*, there is an extra β strand and its orientation is reversed. Interestingly, this part of the protein is not under SCOP domain definition, which means that there is no functional or structural role for this part of the protein. Since this part does not have structural or functional constraints, it is more likely to be subject to mutations and may be influenced by environmental factors of *Bacillus* compared with yeast or *Escherichia coli*. From this, I conclude that the evolution of adenylate kinase would have more likely started from a common ancestor and given rise to *Escherichia coli* and or *Bacillus* and later to yeast protein. Later, *Bacillus* adenylate kinase would have acquired some changes in its protein.

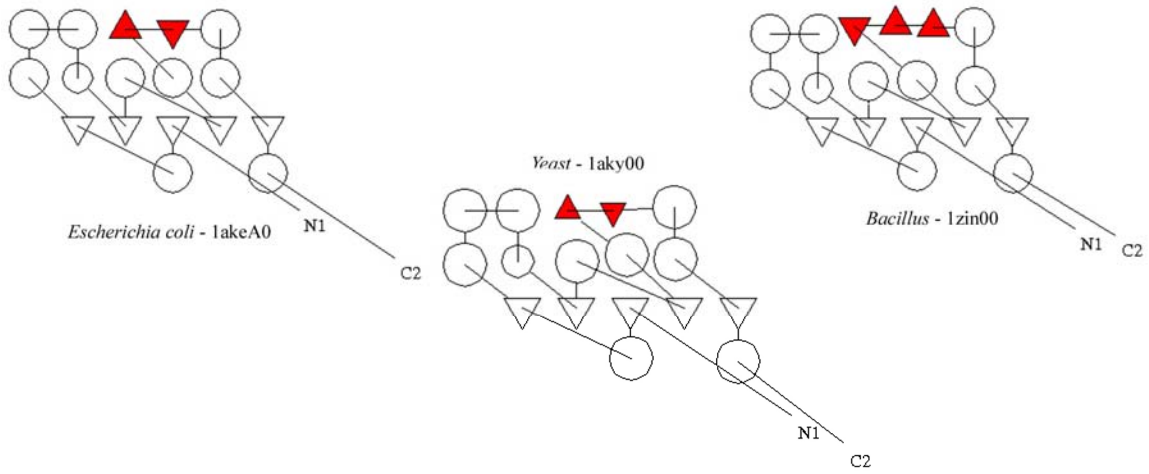

Figure 82. Topology diagram for adenylate kinase

Likewise, from the cytochrome map, two SCOP protein groups form distinct clusters from the rest of the cytochrome members. The overall topology of the cytochrome superfamily members were analyzed using TOPS (Figure 83). Generally, cytochrome c fold has 5 helices. However, some members of *cytochrome* c_{551} group have 6 helices and *cytochrome* c_2 group has 5 helices and 2 β strands except *d3c2c__*, which has only 5 helices. The topology of *cytochrome* c_{552} group (5 helices) remains the same, although its sequence has diverged greatly. However, the domains of this group (*cytochrome* c_{552}) forms close cluster with domains of different cytochrome c protein groups than among itself. It might be one of the typical cases, where orthology/homology cannot be resolved based on sequence identity because an extensive sequence divergence has occurred. However, it can also be argued that *cytochrome* c_{552} proteins were actually formed from convergence of different cytochrome c proteins. But this is highly unlikely to occur given the clear picture of overall divergence of cytochrome proteins and absence of any convergence reports in the cytochrome c fold.

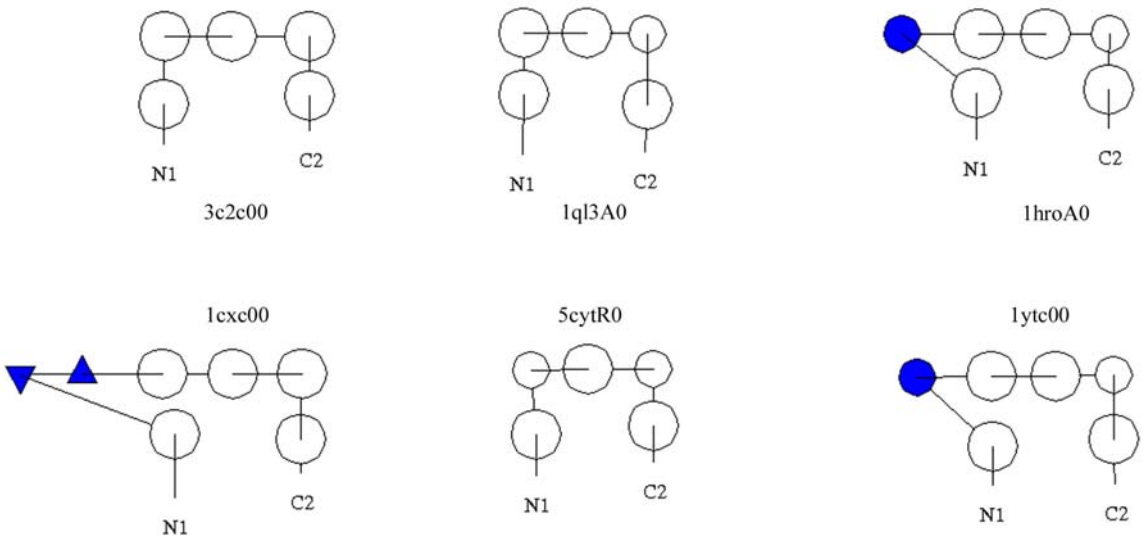

Figure 83. Topology diagram for cytochrome c proteins

Mitochondrial cytochrome c was seen later in the time-scale when compared to bacterial cytochrome c. Given the endosymbiotic hypothesis, it is likely that any bacterial cytochrome c would have given rise to *mitochondrial cytochrome*. Here, it can be seen that *cytochrome* c_2 clustered closely with *mitochondrial cytochrome* (Figure 78). So it is likely that *cytochrome* c_2 would have been the ancestral protein for *mitochondrial cytochrome*. This was confirmed with expertise knowledge of Alexey Murzin. The topology study of these two SCOP protein groups also confirmed this. The general topology of *cytochrome* c_2 and *mitochondrial cytochrome* are 5 helices + 2 β strands and 5 or 6 helices respectively. However, some of the domains of *cytochrome* c_2 (e.g., *d3c2c__*), clustering near to *mitochondrial cytochrome* lack the two β strands, confirming that the earlier forms of *cytochrome* c_2 with β strands, later lost the β strands and have given rise to *mitochondrial cytochrome*.

Thus, cluster maps made with sequence and structural homology is useful in understanding the ancestry of proteins.

B.7 Conclusions

Protein evolution, driven by structural and functional constraints, may leave a trail of homologs. Homologs are identified using sequence comparison methods like BLAST, FASTA, psi-BLAST and ISS. A comparison of ISS with psi-BLAST was made in two

protein superfamilies: cytochrome c and P-loops. The result showed that psi-BLAST detected all the remote homologs identified by ISS in P-loops and only half in cytochrome c superfamily. Although, I cannot generalize using these limited results, it can be said that ISS performs better in some cases than psi-BLAST. The advantage ISS has in some cases might be due to the match score it gains by producing longer alignments around conserved regions of the protein. Intermediate search conducted using structural information revealed that more remote homologs that could not be identified with sequence information alone. So structures might be useful in intermediate search when sequence information is inadequate in detection. From the progressive alignments generated using most of the domains in four SCOP protein groups (*mitochondrial cytochrome*, *cytochrome* c_2, *cytochrome* c_{551} and *cytochrome* c_6), an overall consensus was generated. The highly conserved residues found in the overall consensus are in tandem with the key structural and functional residues needed for the cytochrome c fold (Ptitsyn, 1998). Thus ISS alignments might be useful in understanding highly conserved residues in a protein fold.

Along with sequence information, I used structural comparisons by PrISM to produce a manual cluster map. The cluster map showed a useful representation of the general evolutionary relationships within P-loops and cytochromes. These might be helpful in depicting the relationship between SCOP families, assigning hierarchies to a new protein structure in the existing structural classification and understanding the likely ancestor of a protein. For example, in cytochrome c superfamily, it was shown that the *cytochrome* c_2 protein is likely to be an ancestor for *mitochondrial cytochrome*. The manual process has been automated and can now be used as a tool in exploring evolutionary relationships of any protein family.