CHAPTER 3

Development of a homologue collection pipeline for length-variant proteins

Related reference:


3.1 INTRODUCTION

Evolutionary effects in form of sequence and structural modifications within protein domains, prove to be so strong in some distantly related proteins belonging to protein superfamilies that, despite similar overall structural folds, they may accommodate two-fold length variations [3]. Analysis of the implications of these length variations in the form of insertions/deletions (indels) in protein domain superfamilies on their structure and function was the motivation for Sowdhamini and co-workers to develop algorithms like CUSP (Conserved Units of Protein Structures) [129]. Using CUSP algorithm, 353 multi-membered PASS2 superfamilies was examined earlier by the group to identify structurally conserved or unconserved (mainly indel) regions in an automated manner. Length-deviant and length-rigid superfamilies were differentiated, where the classification was based on the extent of length variation. However, analysis of only the structural entries regime to identify indels, as used in the above study [82,130], may prove to be quite self-limiting and cannot provide a comprehensive picture, thereby failing to model the effects of length variations properly. In this Chapter, this approach is extended to the sequence regime, where homologous sequences are considered in evaluating the effects of length variations. As correctly summed up by Bork & Koonin [1], just the exponential growth of sequence space does not directly cause increase in our understanding about gene and protein functions; but gathering important information like identification, verification of homologues from this vast and redundant knowledgebase is one of the main role-players. Hence, this study is a dedicated effort to search for close and distantly related homologues of protein domain superfamilies, wherein our approach can include homologues with considerable length variations without encountering false positives.

Rapid and heuristic sequence search algorithms (viz. BLAST, PSI-BLAST) [131] are generally sensitive and effective in recognizing protein domains that are distantly related within large sequence databases, but are not well-suited to identify remote homologues of varying lengths. An even more challenging aspect is introduced to distinguish reliable hits from a vast number of putative false positives that could have suboptimal sequence similarities. Here, a data-mining protocol is presented that provides stage-specific filters (described in Methods section) in sequence searches to reliably accumulate remote homologues, which encourages sampling of length variations although with a low false positive rate. 731 superfamilies (only multi-membered superfamilies from PASS2 database of 1961 superfamilies [12]) containing 7439 protein domains is used as queries of starting structural dataset against 9,577,796 sequences in NR database (August 2009 version, NCBI). The task of collecting trusted homologues which has length variability is, therefore, a non-trivial one. And if on average, each protein yields 500-1000
homologues, then total data amounts to nearly 7439x1000 protein hits. Such computationally intense data mining approaches are additionally daunted with the task of specificity i.e. to avoid false positives. As the query proteins within a superfamily are distantly related (<40% sequence identity), they may easily lead to false associations and the ‘hits’ require careful validation, where a hit would be regarded as a true positive only if it belongs to the same superfamily as that of the query. Such computationally intense tasks require a stable pipeline/protocol which can automate the whole process, can be scaled up effortlessly and ultimately provide high signal-to-noise ratio. Realization of remote homologues with vivid length variations after following such protocol will contribute to better understanding of functional variety within protein domain superfamilies.

3.2 MATERIALS AND METHODS

3.2.1 Dataset: Dataset used for initial seed sequence in the search for homologues consisted of structure-based sequence alignments of 1961 multi-membered PASS2 superfamilies [<40% sequence identity between the members] [10,12] from which single-membered superfamilies and two-membered superfamilies were removed. The resultant ~7500 proteins, spread across 731 superfamilies and various length-variant groups, were scaled up to automatically carry out the sequence search protocol as standardized below [Figure 3.1].

3.2.2 Pre-processing query files: Pre-processing of alignment files was carried out and each protein domain was used as query, for which PSI-BLAST (version 2.2.24+) algorithm was initiated against them. Searching for homologues was performed using PSI-BLAST, since it has been proved to be a widely accepted sequence search tool, which is sensitive enough to detect relationships between proteins belonging to the same superfamily [28,132–134]. Usage of PSI-BLAST search for each protein superfamily is corroborated by the fact that database searches often are neither transitive nor symmetric which makes it necessary to perform an exhaustive search iteratively till all the homologues are detected [1]. A particular hit from PSI-BLAST run is considered to be a “true positive”, if it belongs to the same superfamily [according to SCOP database [2]] as the query itself.

3.2.3 Setting parameters and filters for the pipeline: Various parameters used in sequence search or its validation requires to be optimized for the dataset and protocol. The selection of an E-value threshold depends mainly upon the size of database [which in this case is the NCBI-non redundant (NR) database having over 9 million sequences]. Trials with different E-values (10^{-3}, 10^{-6} and 10^{-10}) were carried out in terms of coverage and true positives accumulated. Initially, both E- and H-values were kept constant at 10^{-3} with 20 iterations, as this E-value gave rise to
maximum possible hits. H-values were then varied to observe its effect on position-specific scoring matrix (PSSM) creation and hence its ability to detect remote homologues (H-value is the E-value inclusion threshold for profile-generation in the further rounds of PSI-BLAST). The selection of E-values was performed depending on the coverage analysis (validated true positives out of all hits gathered). Relaxed E-value of $10^{-3}$ was found to yield many false positives, while a stringent E-value $10^{-10}$ yielded many false negatives and did not pick up true homologues. Hence, an optimum was finally fixed at E-value of $10^{-6}$ [Figure 3.2 A, B]. Numbers of iterations were varied and sequence searches of many superfamilies converged after 15th iteration. Hence, the number of iterations was fixed at 20. Though this restriction of 20 iterations may have caused loss of some homologues belonging to ubiquitous proteins (such as RuBisCo small subunit), but it has been established that above a certain number of iterations, alignment accuracy may drop thus affecting the PSSM formed in the process [135].

Gap penalty is a crucial parameter in alignment methods, where its relaxation will allow sequences with high variability in terms of indels to be aligned. Thus, in an effort to accumulate hits with larger length variations, gap penalties of PSI-BLAST were varied at different options (combinations of gap opening penalty, gap extension penalty: 7_2, 9_1, 10_1, 11_1, 12_1) and an optimum value of gap open penalty as 7 and gap extension penalty as 2 on the basis of true positives generated (case study for RmlC cupin-like superfamily [SCOP code: 51182], Figure 3.3) was observed. Similarly, Smith-Waterman [-s] option was kept “ON” to encourage accumulation of short length sub-optimal alignments during sequence searches which allows partial domains and homologues with indels to get aligned. True positive coverage checks, when performed with this option “OFF”, was observed to find lesser hits and less optimal alignments [Figure 3.4].

Apart from the parameter changes in PSI-BLAST, stage-specific filters were introduced as detailed below. To avoid profile drift, which could corrupt the PSSM and hence lead to false connections, the presence of the query was checked amongst the hits at the round of convergence or from the round where query was last seen. Also, since the recognition properties of PSSM change over each iteration (due to inclusion of new hits), there can be cases where high scoring hits in one iteration may be excluded in the next. This may be as a result of the PSSM being influenced by a neighbouring family in the previous iteration and the high scoring hit of interest belonging to that family, i.e. a false positive. Due to filtering of redundant hits which may be a subset/overlap of other segments of the same protein, another stage-specific filter termed as “Merge Filter” was implemented (Algorithm described below). This caused the number of hits to reduce to two-four folds [in cases where hits from all the iterations were taken] [Figure 3.5, 3.6].
Chapter 3: Development of a homologue collection pipeline for length-variant proteins

When performed on one of the superfamilies (superfamily code 51182; E-value=10^{-3}, Gap penalties 7.2) with all hits at all 20 iterations, many redundant entries were obtained. However, if the last round is alone taken, little difference was seen after including the merge filter (please see below). However, the inclusion of merge filter is important to ensure that the hits are unique with all the discontinuous domains joined together and are not subsets of some discontinuous segments or partial domains [Figure 3.7].

3.2.4 Merge Filter Algorithm: Often a single query has significant sequence similarity with two or more regions of the same protein hit. In such cases, taking only one of the regions may cause loss of partial domains or discontinuous domain segments. Here, a “Merge Filter” algorithm is implemented to prevent loss of discontinuous domains and partial domains. The longest length hit (with respect to subject length) in a group of segments [putative homologues] was employed as a standard and each segment was compared with the longest one [Figure 3.5]. Comparison results are presented here for three different cases for which merging treatment had to be done in different ways. The cases and their respective treatments are listed as follows:

a) If both segments are distinct (in terms of start and end positions of both), then they are merged. One checkpoint can be that, if after merging, the length is double that of query length, then they represent “domain duplication” event.

b) If one segment is the subset of a longer segment, then the smaller segment has to be discarded. From the standardization studies, most redundant segments were found to be subsets.

c) If both segments are overlapping, decisions can become subjective and hence a metric, called Extent Of Overlapping (%EOV) is developed.

\[
\text{Extent of overlapping} = \frac{\text{Number of residues overlapping}}{\text{Length of longest segment}} \times 100
\]

If %EOV < 50%, they would be treated as distinct segments and merged eventually. If %EOV > 50%, E-values would be compared and one with the best (lower) E-value would be retained. If the best (lower) E-value one is “not” the longest one, then the second longest of all remaining segments has to be updated as the longest.

PSI-BLAST searches highlighted some small-length hits which spiked the false positive rate by random local alignments. In order to achieve better true positives, a Query Coverage Filter (QCF) of 75% (standardized by earlier studies shown elsewhere) was set upon the initial search
results [Figure 3.8]. Decoding the QCF meant that if the percentage of residues aligned in query (alignment length) with respect to length of query was greater than 75%, then such hits were accumulated. This would ascertain that random fragments do not get aligned and corrupt the search protocol.

3.2.5 Validation of sequence search hits: In order to increase sensitivity (a measure of how many true hits could be obtained from all the sequences in the database) of the search experiments, parameters of PSI-BLAST were varied, but accurate evaluation required validation of the hits obtained; hence, another sequence search method based on HMM was employed for validation [16,136]. A database, (PASS2HMM), was built and employed which consisted of HMM models built from the structure-based sequence alignments obtained from PASS2 (all 1961 superfamilies from PASS2.4). HMMER3 was used to detect domains from the high scoring hits at E-value =10⁻² and ones which did not get associated with the HMM model of the superfamily that contains the query were treated as false positives. Few cases were further checked to confirm if it is indeed a false association by performing fold prediction [137], but this feature was not added in the pipeline, as it was quite time-consuming and difficult to scale up. In order to accumulate maximum remote homologues with large length variations (increasing specificity of process), every hit acquired was double-checked before discarding and this led to different sets of homologues.

A) **Set1:** Hits which have passed 75% QCF and validated by HMMER (on PASS2HMM) form the most trusted homologues set. True positives were the hits which associated with the same superfamily as the query with significant values.

B) **Set2:** Hits which have failed QCF, but have successfully been validated by HMM (on PASS2HMM) have been retained since they might represent partial domains. Since once rejected by QCF, they would have to be scrutinized further regarding their retention of secondary structure, presence of functionally important residues and confirmation if they are fragments or parts of discontinuous domains.

C) **Set3:** Hits which have failed QCF and HMMER (on PASS2HMM) validation, were next validated using their full-length sequences against PASS2HMM and retained if they get associated with the correct superfamily HMM model. These might be cases where insert domains were found or cases of discontinuous domains, which have not been detected by PSI-BLAST earlier, were chanced upon.
Chapter 3: Development of a homologue collection pipeline for length-variant proteins

D) Set4: Hits which have passed QCF but failed to get associated by HMMER (to the correct superfamily of PASS2HMM) were validated against PASS2HMM.1 (1query_HMM) database (PASS2HMM.1 or 1query_HMM: HMM models of each query of all PASS2 superfamilies were individually created and all models appended to create this database). Failure of the hits to get associated to the superfamily level HMM (PASS2HMM) may be due to the fact that HMM model was created using highly dispersed PASS2 members (which have <40% sequence identity with each other); while the single query HMM libraries (PASS2HMM.1) could easily associate with such homologues. Due to the <40% sequence identity cut-off maintained in PASS2, a particular family may have been excluded altogether from the PASS2 superfamily, (eg: Patatin protein in FabD/lysophospholipase-like superfamily (SCOP code: 52151)), and hence the PASS2HMM.1 also was unable to associate with such hits. Thus, single query HMMs were built to validate hits missed by superfamily level HMMs. Thus, this pipeline has stage-specific filters which encourages homologues with high length variations but also discourages randomly aligned false positives by strict validation and stage specific filters.

3.3 RESULTS AND DISCUSSION

Favourable packing arrangements and protein architectures limit the number of possible protein folds, and for this reason large numbers of protein structures might be expected to fall into relatively small numbers of protein fold families. While on a structural level there are 1961 SCOP superfamilies but their sequence counterparts’ needs to be mined from millions of sequences present in NCBI NR database. The homology searches from the sequence space will also lead to selection of sequences which will have an even wider sampling space to collect examples for length variations. The need to mine relevant protein domains with length variant properties from the sequence space can be met by specifically tailoring sequence search protocols.

3.3.1 Effect of parameters optimization, multiple query usage and validation of pipeline

Initially, for a pilot study, one superfamily from each of four different structural classes according to SCOP classification was studied (Alpha class: 47729- IHF-like DNA-binding proteins, Beta class: 51182- RmlC-like cupins, Alpha and Beta class: 56024- Phospholipase D/ nuclease, Alpha by Beta class: 53271- PRTase-like) (Table 3.1). Different PSI-BLAST versions (2.2.21 and 2.2.24+) were tried and due to the higher sensitivity, the latter was chosen to search for hits in NCBI's Non-redundant database (NR database August 2009 version which contains more than 9 million sequences). The different PSI-BLAST parameters which were standardized
[Please refer methods section] and would be followed for the current sequence homologue collection are detailed in Table 3.2.

The full pipeline for collecting homologues [Figure 3.1] was constructed after standardizing the above parameters and instilling many filters to remove redundant and spurious hits. Apart from the PSI-BLAST parameters, filters were applied at other stages, like inclusion of “merge filter” in the pipeline, which proved beneficial in unearthing many cases of putative domain duplication and discontinuous domains which would have been lost if only one region of the hit were to be analysed, as commonly done in other sequence search protocols. Retention of Smith-Waterman filter option ON encouraged the accumulation of short-length sub-optimal alignments with homologues. True positive hits have been annotated as those which have been associated with the superfamily of the query (i.e. HMM model) and rest were assumed to be false positives. False positives analysed were found to possess very low sequence identity (11-30%) with query and also insignificant E-values, which added weight to the assumption that they were truly false positives. To indicate the effectiveness of this pipeline, the true positive bar graphs for different parameters are provided [Figures 3.2-3.8]. The various filters used, were essentially to sieve out homologues which have length variations in them.

PASS2.3 database consists of protein domains arranged at the superfamily level (in correspondence with SCOP 1.73), which are distantly related (<40% sequence identity between members). Since these members are being used as queries for our study, their high sequence dispersion may impose constraints on easy association with sequence homologues. Hence, multiple queries from a single superfamily were used as queries, instead of using one single best representative. Such a step was taken in order to accumulate maximum homologues with varied lengths also [138]. On testing each query (PASS2 structural entry) against PASS2HMM, it was found that most of them picked themselves and few others, but failed to pick all the other members of the superfamily, suggesting the requirement of using multiple queries for search and even for making HMM profile database.

In the vast repertoire of protein domains, several small-sized domains also exist and they sometimes failed to produce any hit at the standardized E-values. Relaxation of E-value to $10^{-3}$ for small query domain (<50 residues) proved beneficial in collection of hits and validation by HMM models further ensured the accumulation of true homologues. Sensitivity of HMM models (both at superfamily level (PASS2HMM) and single query model (PASS2HMM.1) along with their ability of associating hits with the correct superfamily makes it an appropriate tool for validation in this study.
A comparison between our protocol/pipeline and a given E-value cut off (E-value $10^{-5}$ used here) was done. It was found that overall the true positives (%) is higher for one of the superfamilies (SCOP code 51182). A better insight was also given by the false positives, which were quite high when not using our protocol/pipeline [Figure 3.9 A, B]. Hence, this indicates the sensitivity and specificity of the protocol.

Pascarella and Argos observed that length variations were mostly around 5-10 residues long (in 98% of the instances), but they could also be up to two-fold domain length [83]. Hence, it is essential to quantify the length variations wherein the concept of “extent of length variation” which describes length variation as a ratio of the length difference of each member to its average domain size was introduced in Chapter 2, where this metric was further described.

3.3.2 Effect of homologue inclusion on different length-variant groups

After standardizing the parameters for collecting homologues with length variability, the protocol was run for all the 7439 proteins of all 731 multi-membered superfamilies of PASS2 database and the homologues were collated in four different sets. Length variability tests, when performed on the PASS2 superfamilies along with the accumulated homologues (from sets 1 to 3), showed that the percentage length variation in superfamilies has increased. In one instance, homologues with 101% length variation were obtained (in C2H2 and C2HC zinc fingers superfamily (SCOP code: 57667)). Alignments of homologues showing such dramatic increase were created with its query sequence which threw light upon presence of domain duplication events in the homologues. Instances were seen where the inclusion of homologues have changed the length variability character of the superfamily, i.e. becoming a length-deviant from a length-rigid superfamily (data not shown here). Homologues of each query (member of superfamily) at different sequence identity ranges were selected and a phylogenetic tree (Neighbour Joining method) of the full superfamily was created using MEGA [70,139], with a bootstrap value of 1000 (Figure 3.10 for PRTase-like superfamily [SCOP code: 53271] illustrated here. Homologues of each query clustered with them, which is encouraging to justify them as being true homologues. Annotations from NCBI and Gene Ontology were also checked for a few cases taken randomly [140].

Homologue inclusion led to increase in length variation of superfamilies especially in length deviant groups in comparison to the length-rigid ones [Figure 3.11]. On the other hand, length-deviant superfamilies possessed lower sequence identity (distantly related homologues) than length-rigid superfamilies [Figure 3.12].
Due to addition of homologues, the length variation of a superfamily changes and on using the classification technique described in Chapter 2, many superfamilies were found to have switched/shuffled length-variant groups also. Shuffling (switching of superfamilies from one length-variant group to another) has been noted after including the sequence homologues. It was observed that number of length-rigid superfamilies reduces, while number of length-normal superfamilies increases. Though the functional reason for the shuffling is not yet known, but it needs an even closer look into the superfamilies and their homologues. One such example is of Cullin repeat-like superfamily involved in cell cycle, apoptosis which shuffled from being a length-deviant to length-rigid superfamily.

Indels found in length-variant homologues appeared to be directly/indirectly influencing either a functional or structural role, like having diverse functions/multiple repeats (SCOP code: 56024), thermal stability (SCOP code: 47729, 46626), quaternary arrangements (SCOP code: 51182), substrate recognition (SCOP codes: 53271, 53335) and new interaction surfaces (SCOP code: 53067) [82]. Some superfamilies with homologues having varied lengths were checked for their functions in Gene Ontology and a few are explained here.

Rmle-like cupin superfamily (SCOP code: 51182) has mainly auxin-binding protein 1, which apart from binding to auxin also binds to zinc [141]. But other members and their homologues were found to be involved directly/indirectly in cell division during Arabidopsis embryogenesis [142] and positive regulation of DNA endoreduplication (GO: 0032877). While many were hypothetical with no known function, some were even involved in carbohydrate metabolic process by exhibiting mannose-1-phosphate guanylyltransferase/ mannose-6-phosphate isomerase activity [143]. Motif finding tools such as BLOCKS and literature analysis [144] showed a certain motif “LS[I/L]Q[M/V/A]HHH[R/K]” to be conserved in the active site of phosphomannose isomerase (Types I and II) and this overlaps the BLOCK region assigned for Cupin 2 (conserved barrel). The identification of remote homologues consisting of additional subdomains might provide clues to their enhanced or slightly altered biological roles.

ISP transmembrane anchor superfamily (SCOP code: 81502) consisted mainly of ubiquinol-cytochrome-C reductase, apocytochrome F and cytochrome B6, is involved in aerobic respiration, mitochondrial electron transport and ubiquinol-cytochrome-c reductase activity, and also localized in mitochondrion [145,146]. From the GO annotation of this superfamily's members and their homologues, it was evident that they were also similar to Rieske iron-sulfur protein (RIP1) [26] from Saccharomyces cerevisiae [147]. In view of the above examples where length variation leads to functional diversity, it becomes imperative to study such superfamilies
extensively. One of the major steps in that direction needs collecting as many true homologues with length variations.

3.4 CONCLUSION

During sequence searches and evolutionary analysis of protein families, the daunting question is to distinguish between the true positives and false positives. In order to accommodate the length variations without encountering sequence profile drifts, it is also not clear as how much length variations can be tolerated before the accepted ‘hits’ are termed as false positives. When is it deemed fit to associate a hit as a ‘partial domain’ and when should it be deemed as a ‘false positive’? In this study, a dedicated effort has been made to cater to the above questions by organizing a pipeline for collecting length-variant homologues from all multi-membered superfamilies of PASS2 database. Each step of the pipeline has been standardized by varying the default values so as to tailor them to specific needs. The method reliably accumulates remote homologues and further encourages sampling of length variations, along with a low false positive rate. Homologues with length variations were collected for 731 superfamilies from NCBI_NR database. The general outline can be used to collect remote homologues of varying length in an automated mode. Inclusion of sequence entries was observed to lead to an increase in the average length variation in all groups with length-rigid and length-normal groups, thereby showing considerable difference in length variation. Length-deviant superfamilies were observed to have members with extensive range of length variation but comparably lower sequence identity (distantly related) than length-rigid superfamilies. Annotations of homologues for few cases were reviewed and functional diversity, though seen among the superfamily, was not found to be influenced by length variation. Hence, a systematic collection and annotation of homologues has been detailed in the next chapter (Chapter 4), for obtaining insights from the sequence space, about the role of length variation of protein domains on functional and structural aspects. The study has resulted in tracing of length variations across sequence homologues of various domain superfamilies and has been compiled as a knowledgebase of length variant proteins (LenVarDB).
Chapter 3: Development of a homologue collection pipeline for length-variant proteins

FIGURES

Figure 3.1: Pipeline for identification of homologues from the sequence space entries of each superfamily member and their validation using various stage-specific filters.

Figure 3.2: A) Effect of different E-values can be seen on the number of true positives gathered for superfamily 51182 (Rmlc-like cupin superfamily). B) Variation of E-values checked on 4 superfamilies from different structural classes (according to SCOP).

A)
B)

Figure 3.3: Different gap penalties in form of both gap open penalty and extension penalty are varied and their effect on true positive coverage was taken into account while deciding on $7.2$ [Gap open penalty_extension penalty] as the optimal one.

Figure 3.4: Coverage graph for -s T (Smith-Waterman effect) on true positives across 4 superfamilies. Very huge difference in numbers (true positives) in not present but the quality of data collected has improved. a) For four superfamilies from four SCOP structural classes, b) for RmlC-like cupin superfamily (SCOP code: 51182)
Chapter 3: Development of a homologue collection pipeline for length-variant proteins

**Effect of Smith-waterman option in PSI-BLAST upon true positives**

```
<table>
<thead>
<tr>
<th></th>
<th>47729_alpha</th>
<th>S1182_beta</th>
<th>53271_alpha</th>
<th>beta</th>
<th>56024_alpha</th>
<th>AND</th>
<th>beta</th>
</tr>
</thead>
<tbody>
<tr>
<td>-ST</td>
<td>4227</td>
<td>4191</td>
<td>26416</td>
<td>5617</td>
<td>5613</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-SF</td>
<td>4227</td>
<td>4190</td>
<td>26416</td>
<td>5613</td>
<td>5613</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
```

**Effect of Smith-Waterman option on true positives of RmIC cupin-like superfamily**

```
<table>
<thead>
<tr>
<th></th>
<th>d11ha-</th>
<th>d1p4ta-</th>
<th>d1y70a-</th>
<th>d1y2a-</th>
<th>dlyhfa1</th>
<th>dlyuda1</th>
<th>dimps1</th>
</tr>
</thead>
<tbody>
<tr>
<td>-ST</td>
<td>63</td>
<td>1010</td>
<td>1030</td>
<td>999</td>
<td>256</td>
<td>418</td>
<td>414</td>
</tr>
<tr>
<td>-SF</td>
<td>63</td>
<td>1011</td>
<td>1029</td>
<td>995</td>
<td>258</td>
<td>418</td>
<td>415</td>
</tr>
</tbody>
</table>
Figure 3.5: Cartoon illustration of algorithm of Merge/Redundancy filter explained with different case studies.

Figure 3.6: Effect of Merge/Redundancy filter on true positive coverage shown for superfamily 51182 (Rmlc-like cupin). (Hits collected by running PSI-BLAST with E-value 10⁻³.)
Figure 3.7: Effect of merge/redundancy filter on true positive coverage of 4 superfamilies from different structural classes (47729, 51182, 56024, and 53271) and also for 51182 at E-value $10^{-3}$, once with all hits from all iterations and other time with only hits from last round of convergence.

Figure 3.8: Coverage graph for effect of QCF on true positives across 4 superfamilies
Figure 3.9: Comparison of effectiveness between results after default usage [E-value= $10^{-5}$] and results after usage of standardized pipeline parameters, in terms of, a) true positives, and b) false positives

a)

![True positives comparison between pipeline parameters and default (%)](image)

b) 

![False positives comparison between pipeline parameters and default (%)](image)
Figure 3.10: Phylogenetic tree of homologues with their queries for superfamily [PRTase-like superfamily [SCOP: 53271]]. Neighbour-joining method has been chosen for clustering and one member from each class has been chosen. Each query’s homologues have clustered with them, thus proving further their justification as being true homologues.
Figure 3.11: A) Increase in average length variation across length deviant, rigid and normal groups after inclusion of sequence homologues to the structural set of superfamilies. B) Amount of increase in length variation (marked by red bars) by inclusion of homologues as compared to length variation (marked by blue bars) by structural entries of the superfamilies.
Figure 3.12: Distribution of sequence identity (upper panel) and length variation (lower panel) across length deviant (A, C) and length rigid (B, D) superfamilies. Boxplots are useful for comparing distributions between several groups or sets of data. The spacings between the different parts of the box help indicate the degree of dispersion (spread) and skewness in the data, and identify outliers. Outliers are extreme values in comparison to the rest of the data. They can be divided into "mild" and "extreme" outliers. Mild outliers are any score more than 1.5*IQR from the rest of the scores. Extreme outliers are any score more than 3*IQR from the rest of the scores. IQR = Interquartile range, which is the middle 50% of the scores.
### TABLES

Table 3.1: Details of superfamilies used in the pilot study

<table>
<thead>
<tr>
<th>Structural Class</th>
<th>SCOP code</th>
<th>Superfamily name</th>
<th>Number of Members</th>
<th>Average domain size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alpha</td>
<td>47729</td>
<td>IHF-like DNA-binding proteins</td>
<td>5</td>
<td>82</td>
</tr>
<tr>
<td>Beta</td>
<td>51182</td>
<td>RmlC-like cupins</td>
<td>7</td>
<td>127</td>
</tr>
<tr>
<td>Alpha and Beta</td>
<td>56024</td>
<td>Phospholipase D/nuclease</td>
<td>10</td>
<td>205</td>
</tr>
<tr>
<td>Alpha by Beta</td>
<td>53271</td>
<td>PRTase-like</td>
<td>26</td>
<td>184</td>
</tr>
</tbody>
</table>

Table 3.2: Parameters standardized for sequence search pipeline for accumulating homologues with higher length variation.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Values/Remarks</th>
</tr>
</thead>
<tbody>
<tr>
<td>E-value</td>
<td>$10^{-6}$</td>
</tr>
<tr>
<td>H-value (for inclusion in PSSM)</td>
<td>$10^{-6}$</td>
</tr>
<tr>
<td>Iterations</td>
<td>20 (hits to be taken only from round of convergence or last round)</td>
</tr>
<tr>
<td>Locally optimal Smith Waterman alignments option</td>
<td>Yes</td>
</tr>
<tr>
<td>Gap penalty [gap open: gap extension]</td>
<td>7:2, 11:1 (for some superfamilies)</td>
</tr>
<tr>
<td>Query coverage filter</td>
<td>ON</td>
</tr>
<tr>
<td>Merge filter</td>
<td>ON</td>
</tr>
</tbody>
</table>