Methods for improving and characterizing gene families

1 Abstract

2 Specific aims

1. Developing and assessing pair classification based method for detecting under-clustered gene families

2. Developing and assessing tree based and domain based methods for detecting over-clustered gene families

3. Workflow for obtaining consensus families from multiple family sets

4. Biological characterization of gene families in legumes

3 Introduction

3.1 Gene families

Gene families or orthologous groups, with respect to a given set of species, are clusters of genes that have originated from a single ancestral gene present in the common ancestor of the given set of species. Therefore, gene families contain genes that have diverged due to speciation or duplication after the divergence node of earliest diverging species. Accordingly, genes within the families are either orthologs or in-paralogs (duplication after the common ancestral node). (Fitch 1970, 2000; Sonnhammer and Koonin, 2002). Gene families are mainly used to infer orthology among the given set of species.

Clustering genes into families helps in capturing evolutionary information from multiple species for studying the divergence or conservation of particular biological functions, with respect to given set of species. Gene families can also be used for identification of candidates drug/vaccine development in case of parasites like Plasmodium Falciparum which causes malaria in humans (Gardner et al. 2002; Kissinger et al. 2002; Bahl et al. 2003). Gene families have also been used for annotation of newly sequenced genomes and for anti-bacterial drug development by cross-referencing functional information from multiple species

(Tatusov et al. 1997; Galperin and Koonin 1999; Natale et al. 2000a,b; Forterre 2002)

3.2 Improving and characterizing gene families

The Clusters of Orthologous Groups (COG) database is one of the earliest methods for building gene families (Tatusov et al. 2000, 2001) where orthologous groups are identified based on reciprocal best hits from all-against-all BLAST searches of complete proteomes among three distinct phylogenetic lineages. Existing family building algorithms can be divided into 2 main categories viz: BLAST/alignment statistics based clustering approaches and tree based approaches. The BLAST based clustering approaches either operate on individual pairs of species and then expand to multiple species, example - MultiParanoid (Inparanoid), OMA (Alexeyenko et. al 2006; Remm et. al. 2001 ;Roth et. al. 2008; Altenhoff et. al. 2011), or attempt to infer families directly from multiple species with the common ancestor of given set of species as a reference like OrthoMCL, OrthoFinder (Li et. al. 2003; Emms et.al. 2015). Tree-based methods compare the gene trees with the species trees using tree reconciliation methods for building gene families (Li et. al.; Van der Heijden et. al.; Wapinski et. al.; Hubbard et. al.; Dehal et. al.).

Although, majority of families produced by existing methods can be expected to be consistent and true, many non-optimal families could be produced due to under-clustering or over-clustering. Under-clustered families are unusually small clusters that may have been fragmented due to low recall (missing sequences). Over-clustered families are unusually large sequence clusters that may been produced due to merging of two or more closely related families. Also, a significant number of sequences can remain un-clustered and could be wrongly considered as singletons or orphan genes.

Therefore, it is important to post-analyze and assign confidence scores to individual families obtained using the existing methods based on family accuracy. Here, I propose a suite of methods and tools for improving under-clustered/over-clustered gene families, and for characterizing true and consistent families.

3.3 Gene families in legumes

The legume family (Fabaceae) consists of 750 genera and 19,500 species of flowering plants making the family the third largest family of flowering plants (Lewis et al. 2005). The family is broadly divided into four major lineages viz. Papilionoideae, Mimosoideae-Cassiinae-Caesalpinieae (MCC), Detarieae, Cercideae out of which the Papilionoideae is the largest. The Papilionoid sub-family includes species like Arachis hypogaea (cultivated peanut), Arachis duranensis and Arachis ipaensis (wild peanut varieties), Cajanus cajan (pigeonpea), Cicer arietinum (chickpea), Glycine max (soybean), Lotus japonicus (Lotus), Lupinus angustifolius (narrow-leafed lupin), Medicago truncatula (barrel medic), Phaseolus vulgaris (common bean), Trifolium pratense (red clover), Vigna angularis (adzuki, bean), Vigna radiata (mungbean) and Vigna unguiculata (Cowpea).

An ancient Whole Genome Duplication (WGD) has occurred in common ancestor of the Papilionoid sub-family around 55 Ma. (Blanc and Wolfe 2004; Schlueter et al. 2004; Pfeil et al. 2005; Cannon et al. 2006; Bertioli et al. 2009; Cannon et. al. 2014). In addition, the Glycine and Lupin lineages have also undergone independent lineage specific WGDs (Cannon et. 2014; Kroc et. al 2014)

Approximately 17k-18k gene families were obtained using a Ks-based (synonymous substitutions per site) family building procedure from the Papilionoid species proteomes. Additionally, a different set of 20k families was also obtained from Papilionoid species using the OthroFinder method (Emms et.al. 2015) for comparison.

4 Objective 1: Pair classification based family scoring

This scoring method leverages the basic evolutionary property of gene families - all the sequences in the family must have diverged at/after the separation of earliest diverging species and none of the sequences outside the family have diverged in the family. Accordingly, for a given family, this method attempts to separate pairs from the family, from pairs between the family sequences and non-family sequences. The separation statistics tell how well a given family of sequences can be circumscribed from closest non-family sequences, which can provide an indication of family completeness.

4.1 Methods

4.1.1 Family searching and collecting closest non-family sequences

All the sequences from any given family are searched against the combined database of all proteomes from all the species under study (the sequence space). With each family sequence as query, all the hits that match the query better than the worst matching family hit (other than the query) are collected into a list. A combined list of hits, using each family sequence as a query, is obtained which is expected to contain both family and non-family sequences (figure: 1).

The phmmer program from HMMER package (Eddy, S., 1992) is used for searching families against the sequence space database. The ranking of hits for finding the worst matching family hit is the same ranking given by the phmmer search results.

4.1.2 Forming family and non-family pairs

The combined list of family and non-family sequences obtained in the previous step is used to form family and non-family pairs (figure: 2). The family pairs are those that form exclusively between the original family members and the non-family pairs are those that form between the family and non-family members. The family pairs are labeled as positive pairs and the non-family pairs were labeled as the negative pairs.

4.1.3 Model building and assessing classification performance

HMM based pair classification models are built to classify the family and non-family pairs, and 10 iterations of repeated test-train split strategy are used to assess the classification performance (figure: 3). For each iteration, the set of family+non-family pairs are randomly split into training and test split. The pair classification model is trained on the training split and tested on the unseen test split. To train the HMM model on positive pairs in the training split, consensus sequences of the positive pairs are used. To test the model on pairs in the test split, individual sequences of the pairs are first aligned to the trained HMM
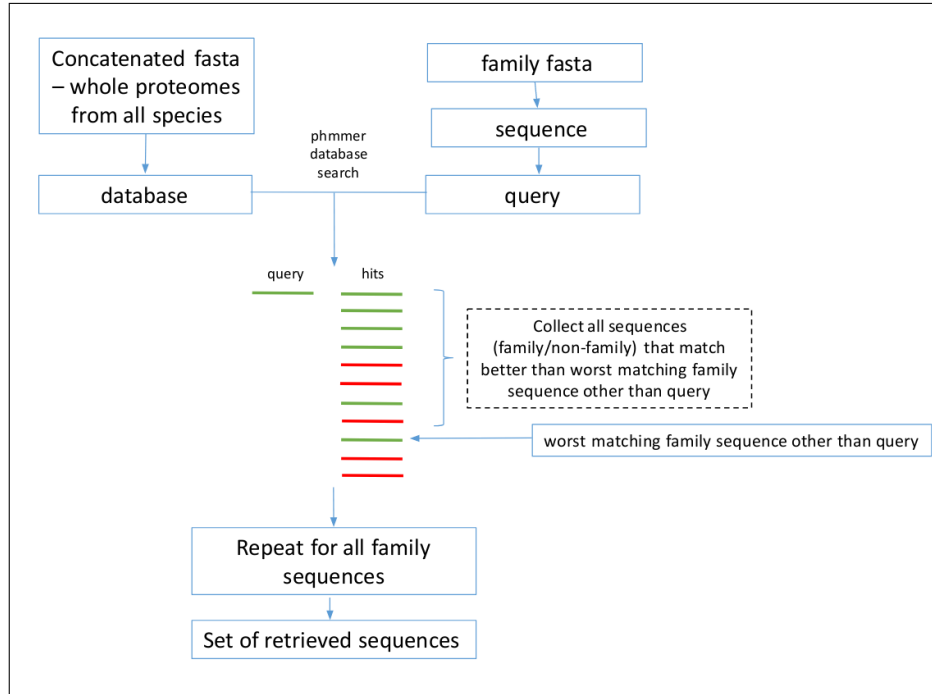
4

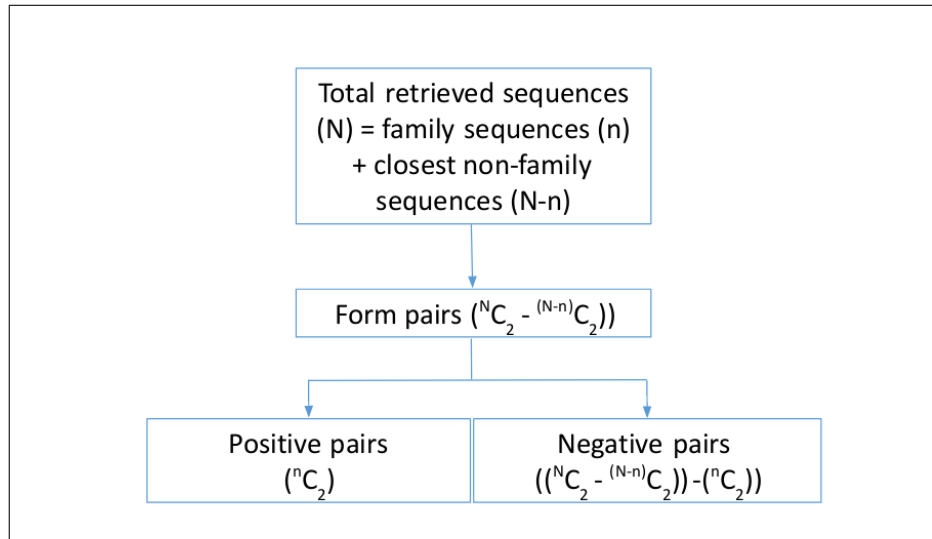Figure 1: Collecting closest non-family sequences

Figure 2: Forming positive and negative pairs from family and non-family sequences

model and alignment scores for both the sequences are obtained. The test pair is predicted as a positive pair if both the alignment scores are greater than or equal to a specified score cutoff.

The Precision-Recall (PR) curve for the positive class is obtained using the combined predictions from all the 10 test-train split iterations. The test pair alignment scores from all the iterations are consolidated (figure: 4) and used for calculating precision (TP/TP+FP), recall (TP/TP+FN) and F1-score (harmonic mean of precision and recall) values for specified score cutoffs. Different precision and recall values are obtained for a range of score cutoffs starting from most stringent to least stringent. The precision values are plotted against the recall values to obtain the PR-curve (Davis et. al 2006). The area under the PR-curve (PR-AUC) is calculated using the trapezoidal rule. Other classification metrics such as the precision and recall values observed at the best F1-score and the score cutoff which gives the best F1-score are also reported. The lowest score observed for the positive class is also reported as the lowest score cutoff with respect to the given family.

An example of PR-curve plot for a hypothetical gene family is shown in figure: 5. Each point on the curve corresponds to an alignment score cutoff with score cutoffs decreasing from left to right. The highest score cutoffs on the left produce classifications with highest precision (low FP) but lowest recall (high FN). Conversely, lowest score cutoffs on the right produce classifications with lowest precision but highest recall. An F1-score can be calculated for each point (precision, recall) on the curve as the harmonic mean of precision and the corresponding recall value. The point on the curve with the highest F1-score is point where optimal values of precision and recall exist. This point represents the best classification performance for the family and the corresponding alignment score is the score cutoff that gives the best classification performance for the family.

4.1.4 Predicting missing sequences

Candidate missing sequences for every family are predicted using the non-family pairs. All the non-family pairs where the scores for both the sequences of a pair are greater than the best/lowest score cutoff are re-classified as positive pairs. Unique sequences within these reclassified positive pairs are predicted and reported as candidate missing sequences for the family.

4.2 Preliminary results

The pair classification based scoring was tested on 4796 yeast families from the Yeast Gene Order Browser (YGOB) database (Byrne et. al. 2005). Since the YGOB families are built through manual curation using synteny based evidence, they were assumed to be true. The distribution of the PR-AUCs for all the 4796 yeast families was obtained (figure: 6) to check if the quality assessment workflow is assigning high quality scores to all the true families. As expected the distribution of PR-AUCs is highly skewed towards PR-AUC value = 1.0 with
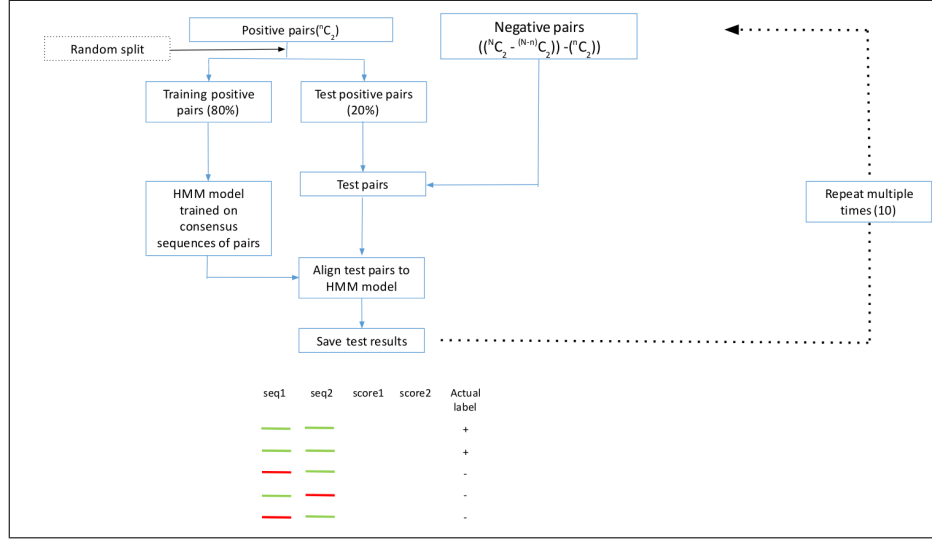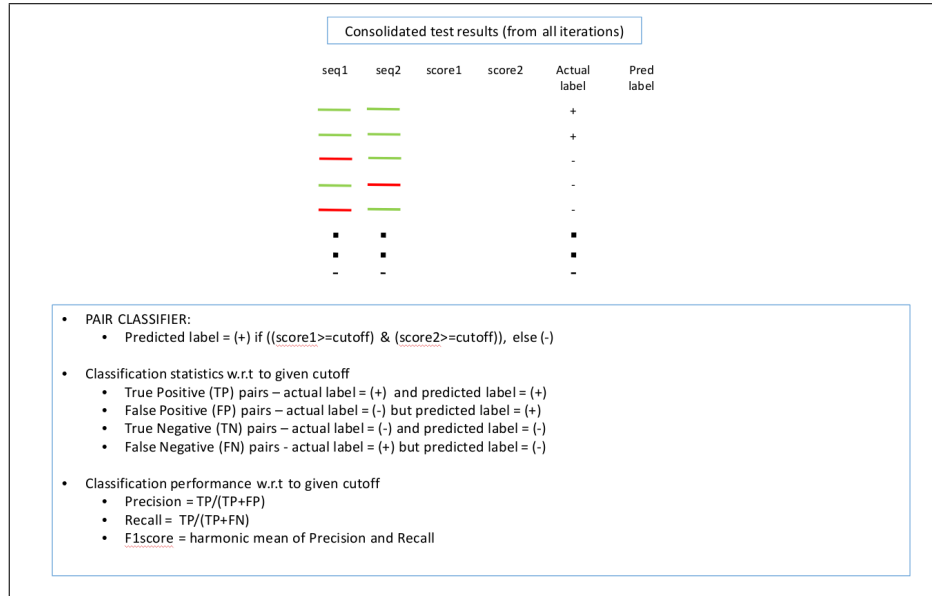
Figure 3: Training and model evaluation
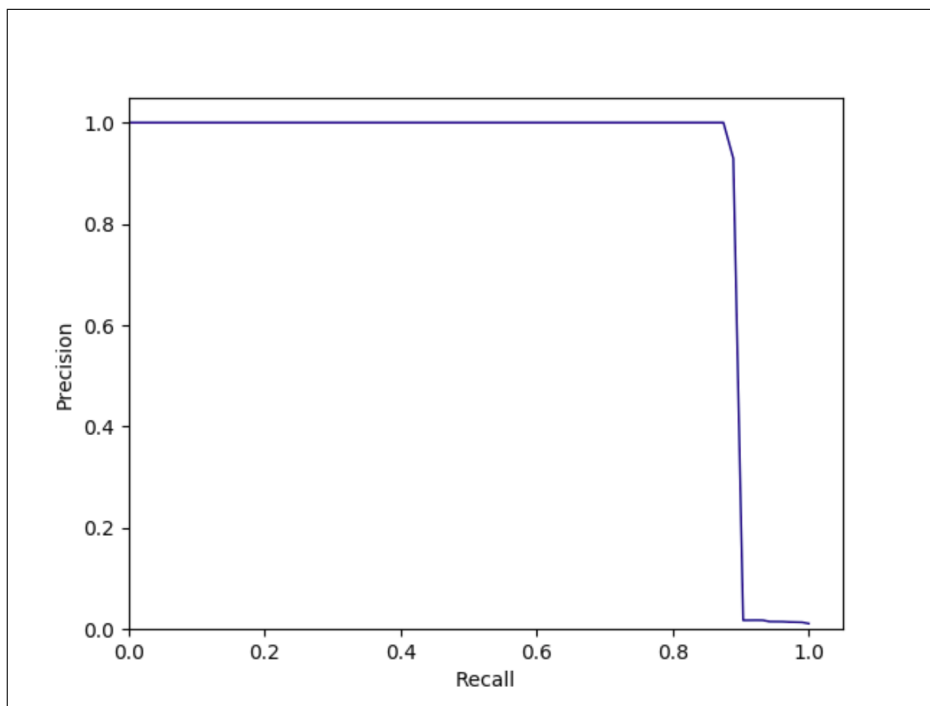

Figure 4: Test pair classification

Figure 5: Example of PR-curve obtained during training

92% of family classifiers having values ≥ 0.75. This shows that the method correctly recognizes good quality families and assigns high quality scores to them.

Even though the quality assessment method is proven to work on good/true families it is important to study the behaviour of the method on non-optimal families. To check the behaviour of the method on incomplete families, the true yeast families were artificially manipulated so that all the families are randomly missing 20% of family sequences.

The histogram in figure: 7 shows the distribution of PR-AUC values for 4796 artificially manipulated families where every family is missing 20% of their sequences. A scatter-plot (figure: 8) of true families PR-AUCs v/s PR-AUCs after removing family sequences shows the drop in the PR-AUCs after removing family sequences from true families. The distribution indicates that the overall quality of the families has dropped significantly which shows ability of method to detect incomplete families.

For each manipulated family with missing sequences, we also predicted the missing sequences back using lowest score cutoff obtained during training. The predicted missing sequences were compared to the true missing sequences and, precision and recall values were calculated to study the accuracy of predicted missing sequences. Precision was calculated as (TP/(TP+FP)) and recall was
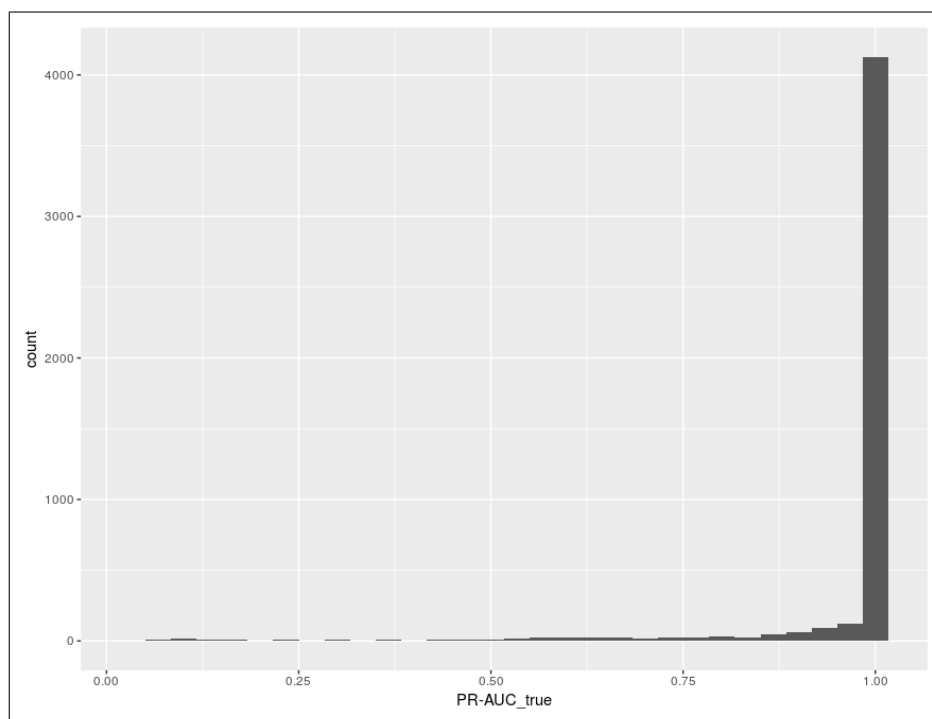
8

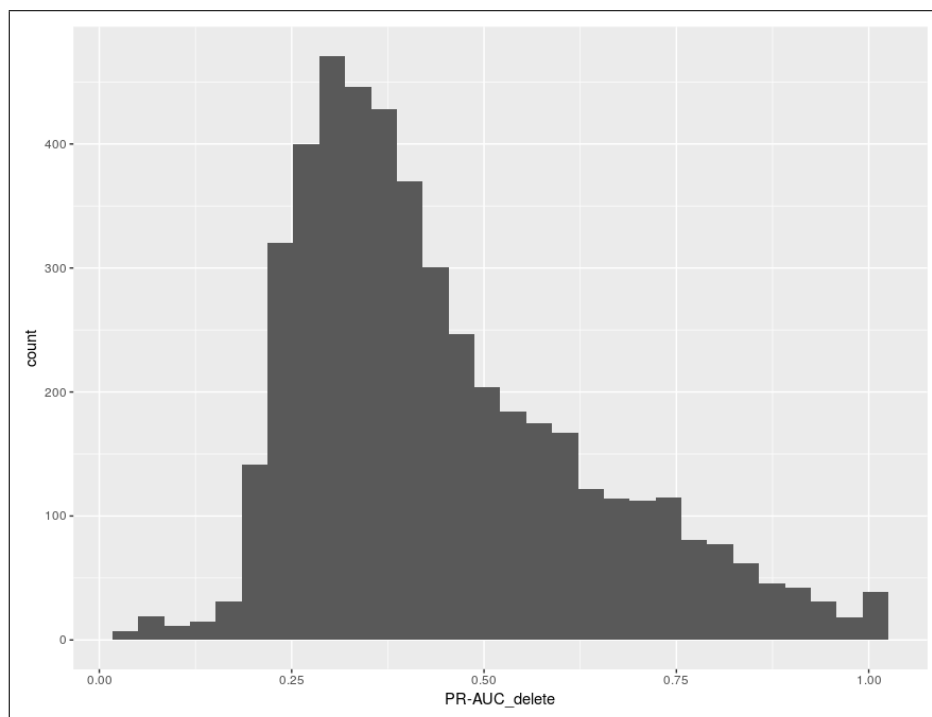Figure 6: Distribution of PR-AUC values for true YGOB families

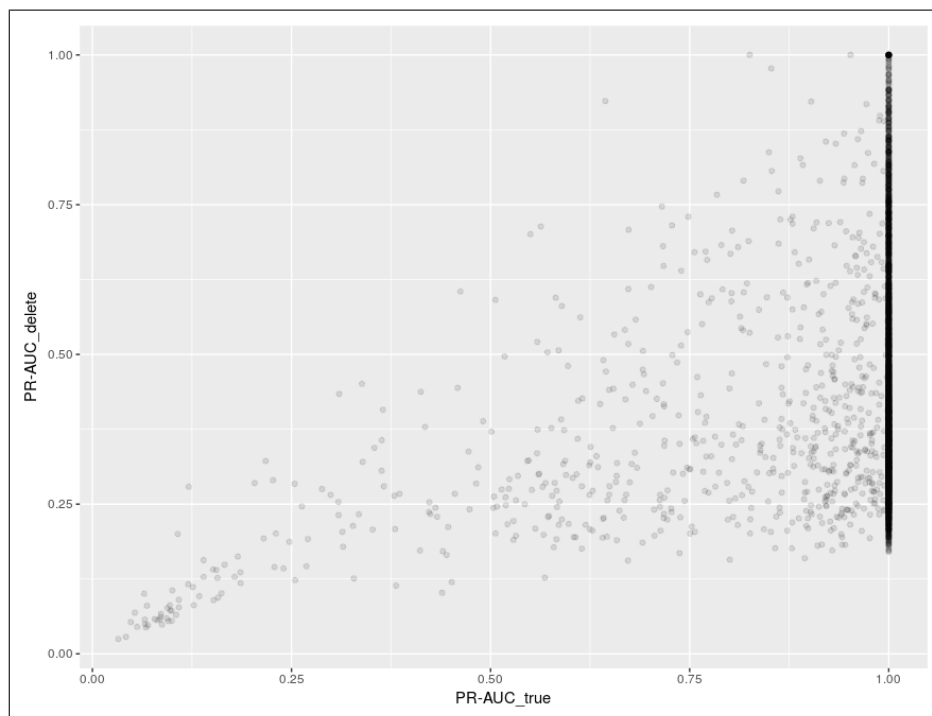Figure 7: Distribution of PR-AUC values for families with 20% of sequences deleted

Figure 8: Plot of PR-AUCs obtained from true families v/s PR-AUCs after deleting 20% of sequences

calculated as (TP/(TP+FN)) where TP are those predicted missing sequences that were truly missing from the family, FP are those that were predicted as missing but do not actually belong the family and FN are those that are truly missing but were not predicted as missing by the method. Out of 4796, the prediction performance for missing sequences was good (precision ≥ 0.75 and recall ≥ 0.75) for 3760 (78.4%) families.

4.3 Signficance

This family scoring method assumes the given family to be under-clustered i.e. majority of sequences have diverged into the family and attempts to build an HMM-based classifier that classifies pairs of sequences formed within the family from the pairs formed between family and the closest non-family sequences. The classification performance is mainly given by the PR-AUC value which shows how well the given family is separable from the closest families. This scoring method can be used in two ways viz. to characterize true families and to detect missing sequences for under-clustered families. For characterizing true families, the PR-curve and the PR-AUC can be used to see how well-behaved the family is in terms of its evolution. The best score cutoff obtained during training can be used to add sequences to the families from newly sequenced species. For under-clustered families, the PR-AUC is expected to be low as the precision value drops due to greater number of FP pairs (sequences trying to get in the family). The missing family sequences can be predicted using the best or the lowest score cutoffs depending upon preference of high precision or high recall for prediction. If precision is preferred over recall then the best score cutoff can be used and if recall is preferred over precision then the lowest score cutoff can be used for predicting missing sequences.

4.4 Future work

4.4.1 Merging small legume families

This scoring method will be used to inspect and merge small legume families into larger families. This will help in reducing the number of under-clustered families and generating a set of high-quality legume families.

4.4.2 Solving scalability issues

Currently, the pair classification based method is not fast enough to be efficiently applied on all legume the families. To solve these scalability issues, multi-threading can be implemented to execute the test-train split iterations parallelly.

5 Objective 2: Methods for detecting over-clustered families

5.1 Tree based family scoring

Family trees are largely expected to follow the pattern of evolution from the species tree. This equivalence between gene trees and species has been used previously in multiple methods like TreeFam, LOFT, SYNERGY, ENSEMBLE 2007, PhIGs (Li et. al.; Van der Heijden et. al.; Wapinski et. al.; Hubbard et. al.; Dehal et. al.) for building gene families and detection of orthologs and paralogs from the family trees. Animal gene families in TreeFam are the group of genes evolved from a single gene in last common of all animals. Genes in one family are identified on the basis that they are phylogenetically separated by one or more genes from the outgroup species (S. cerevisiae, S pombe. or A. thaliana)

Since gene families contain sequences that have diverged at or after the divergence of the earliest diverging species under study, genes within families have also diverged after the divergence of outgroup species. This evolutionary property of gene family can be used to score family trees by calculating the proportion of ingroup sequence pairs that appear to have diverged after the divergence of outgroup sequences in the corresponding evolutionary trees of the families.

5.1.1 Methods

For a given rooted family tree, each pair of ingroup sequences, found within the tree, is labeled as True Positive (TP) or False Positive (FP) depending upon whether the pair appears to have diverged after or before the divergence of one or more outgroup sequences. A precision score for the family is calculated as TP/(TP+FP) which gives the proportion of ingroup pairs diverging after the outgroup separation in the family tree.

To check the divergence of any ingroup sequence pair in the tree, the MRCA of the pair is obtained. Then, all sequences corresponding to leaf nodes under this MRCA are collected. If this set of sequences contains one or more outgroup sequences, the corresponding ingroup pair is labelled as FP, else it is labeled as TP.

5.1.2 Preliminary results

This scoring method was applied on 18543 legume families. The histogram (figure: 9) shows the distribution of scores across the legume families. Most of the families have higher precision scores ($\geq$0.75).

A scatter-plot (figure: 10) of tree-based precision scores v/s number of ingroup sequences (fig) was also obtained to study the relationship between the precision scores and family size. As expected, most of families containing 15-45

ingroup sequences have high precision scores. A considerable number of families, containing approximately 40 ingroup sequences, have tree precision scores around 0.5. These are the families where approximately half of ingroup pairs are diverging after the divergence of one or more outgroups sequences. These families are prime candidates of over-clustered families. Also, large families with very low precision scores (majority of ingroup pairs have diverged before the outgroup divergence) may indicate problems with tree rooting.
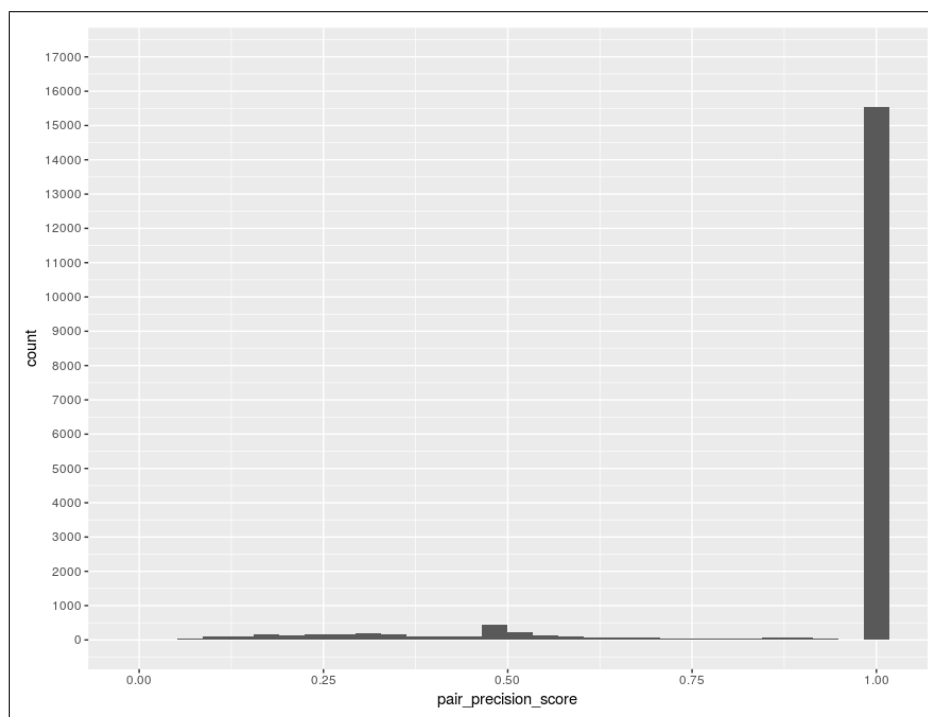

Figure 9: Distribution of tree based precision scores of lgf5 families

5.1.3 Significance

This simple tree-based scoring method uses basic evolutionary properties of gene families to detect if all the sequences clustered in the family truly belong in the family using the rooted evolutionary family tree. As seen from the preliminary results, this scoring method can be used to detect over-clustered families and family trees with rooting problems.

5.1.4 Future work

Over-clustered legume families detected using this method will be corrected to obtain a set of high-quality legume families. Also, large families with low
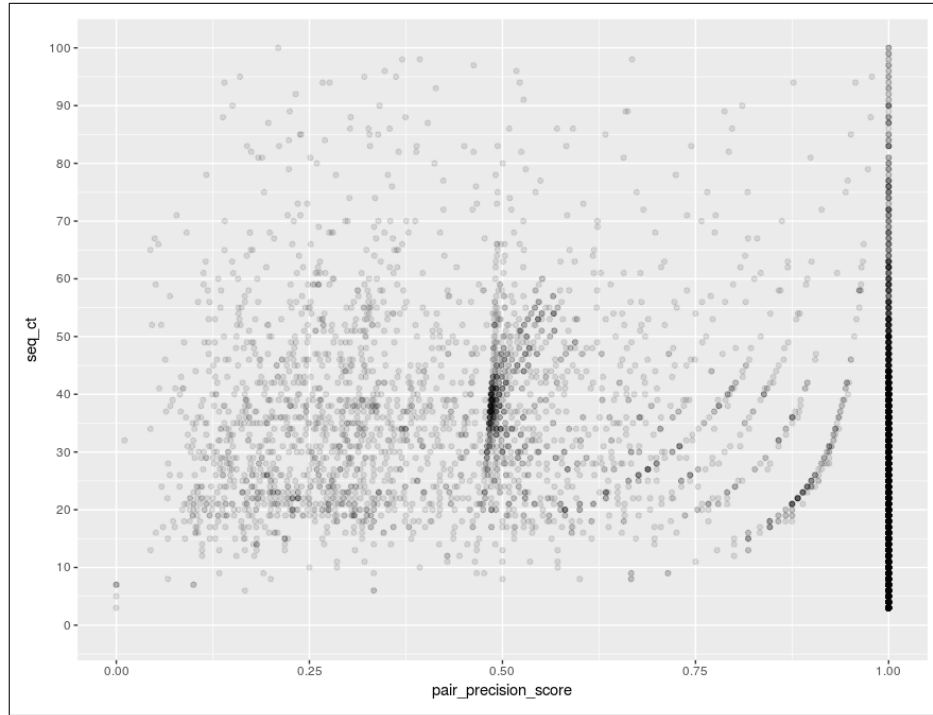
Figure 10: Plot of tree based precision scores v/s ingroup sequence count for lgf5 families

precision scores will be re-rooted to maximize the precision scores.

The merged families obtained using the pair based scoring will also be analyzed using the tree based scoring to resolve any over-clustered families produced due to merging.

An R-package will be developed where users can analyze and score family trees with respect to a given set of outgroups. The R-package will also be able to directly extract ingroup clades from over-clustered families.

15

5.2 Domain composition based family scoring

Protein domains are sections of protein sequences that can fold and function independently. Consequently, multidomain proteins can evolve in a modular fashion through domain deletions/insertions/duplication in addition to sequence-based evolution. Databases like Pfam, PROSITE, SMART, CDD (Finn et. al.; Falquet et. al.; Letunic et. al.; Marchler-Bauer et. al.) define and store domains detected in large number of protein sequences. Domain compositions/content have been previously used to detect sequence homology with high accuracies. (Song et. al, Bitard-Feildel et. al). However, these were well separated gene families with distinctly different domain compositions (E.g. families separated at metazoan ancestor)

Closely related sequences from the same family can be expected to have fairly same domain types and domain content. Accordingly, this method assigns scores to given families based on number of domains shared between the sequences of the family. This scoring scheme will assign higher scores to family where most of the sequences have similar domain compositions. Good and well-conserved families are expected to have higher domain composition scores. Conversely, over-clustered families containing diverse sequences are expected to have lower scores due to less conservation of domain compositions across all the sequences.

5.2.1 Methods

5.2.1.1 Domain composition based jaccard score

For a given family, Pfam-A domains are assigned to all the sequences in the family using the pfamscan program (Mistry et. al.) The domains are assigned to sequences in order of their start coordinates in the sequences. Then, for each pair of sequences in family, domain composition based jaccard score is calculated using the function: $(n1 \cup n2)/(n1 \cap n2)$, where n1 and n2 are the sets unique domains from sequence 1 and 2 of the pair, respectively. Jaccard score for the family is calculated as mean of all pairwise jaccard scores from the family.

5.2.1.2 Domain feature vector based cosine score

Here, the domain compositions of sequences within a given family are compared using their domain feature vectors. This is new scoring method for comparing domain compositions of two sequences. For example, consider 2 sequences X and Y from the same family having following domain order along the sequence, X: {A,B,B,C} and Y: {A,A,B,D}. The duplicate domains in both the sequences are assigned unique ids to distinguish them from each other, X: {A, B, B-2, C} and Y:{A, A-2, B, D}. The domain content universe for both the sequences is (A, A-2, B, B-2, C, D). Accordingly, the domain feature vector for both sequences is X: $(x_1, 0, x_2, x_3, x_4, 0)$ and Y: $(y_1, y_2, y_3, 0, 0, y_4)$, where x_i and y_i are the alignment scores for the corresponding domain HMMs aligning against the sequences X and Y, respectively. Cosine similarities are calculated between all pairs of sequences using their domain feature vectors and the domain composition score for the family is calculated as the mean of all pairwise cosine scores.
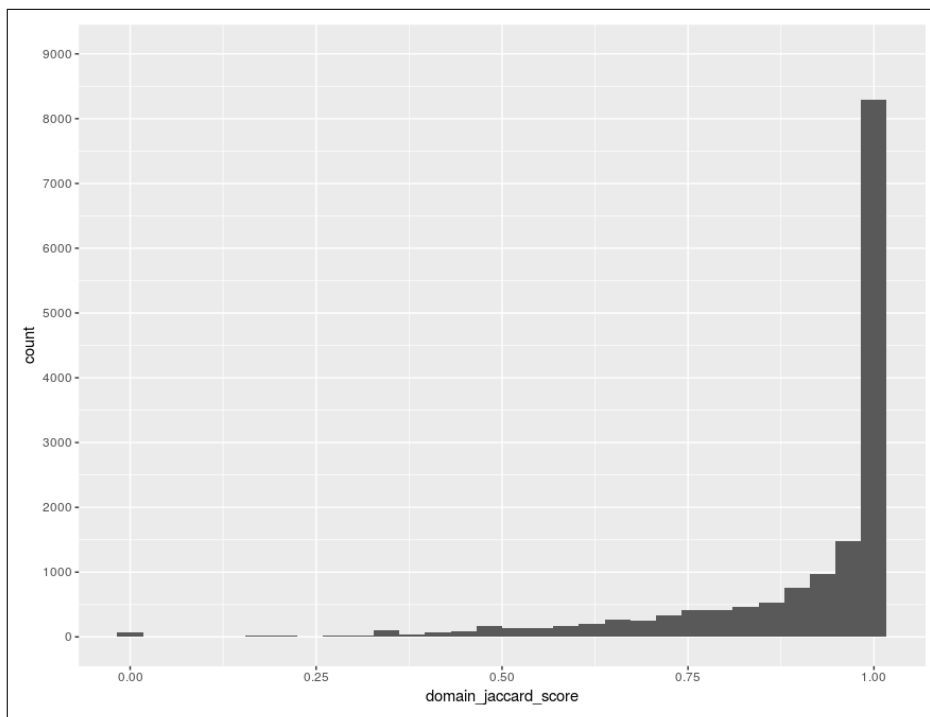
Figure 11: Distribution of domain composition based jaccard scores for the lgf5 families

Pfam clusters closely related domains in Pfam-A into clans. For sequences within the same family, closely related domains could be detected interchangeably, for same positions, in closely related family sequences. To avoid this, clan-ids of the domains are used instead of domain-ids for calculating the domain feature vector based cosine scores.

5.2.2 Preliminary results

The domain composition based jaccard and cosine scoring were applied on the existing legume families to study the relationships between different scores and the ingroup sequence count. The score distributions for domain-based jaccard scores can be seen from figure: 11

To study the relationship between the domain jaccard score and family size, scatter-plot of jaccard score v/s number of ingroup sequences in the family was obtained (figure: 12). The jaccard scores are high for majority of families with number of ingroup sequences between 15-45.

The distribution of domain feature based cosine scores was also obtained (figure: 13). Most of the families have high scores indicating that for majority of the families, the domain compositions are fairly conserved.
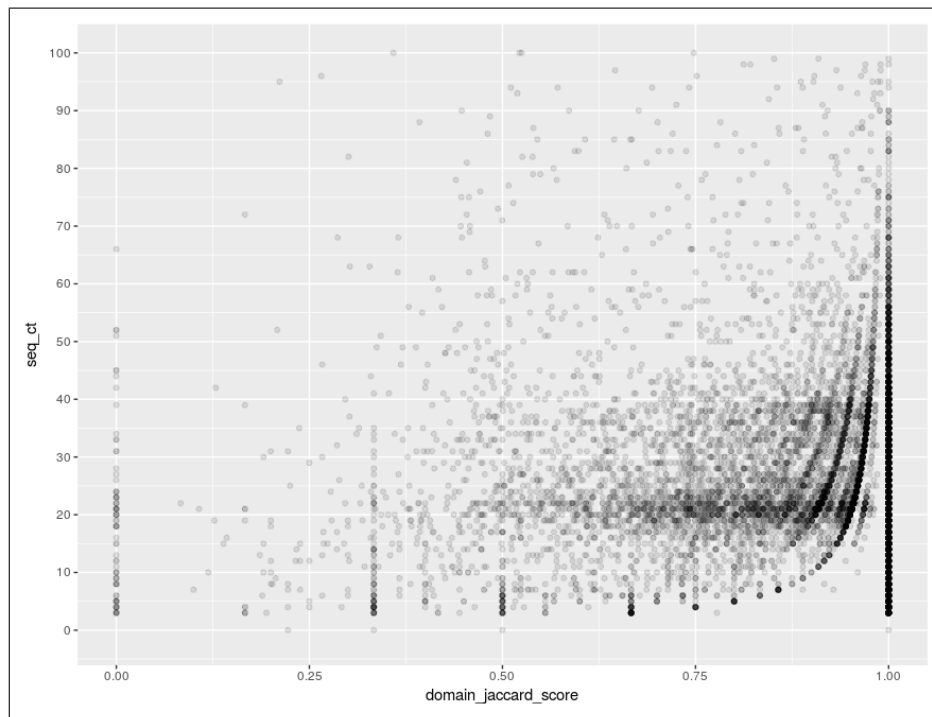
Figure 12: Plot of domain composition based jaccard score v/s ingroup sequence counts for lgf5 families
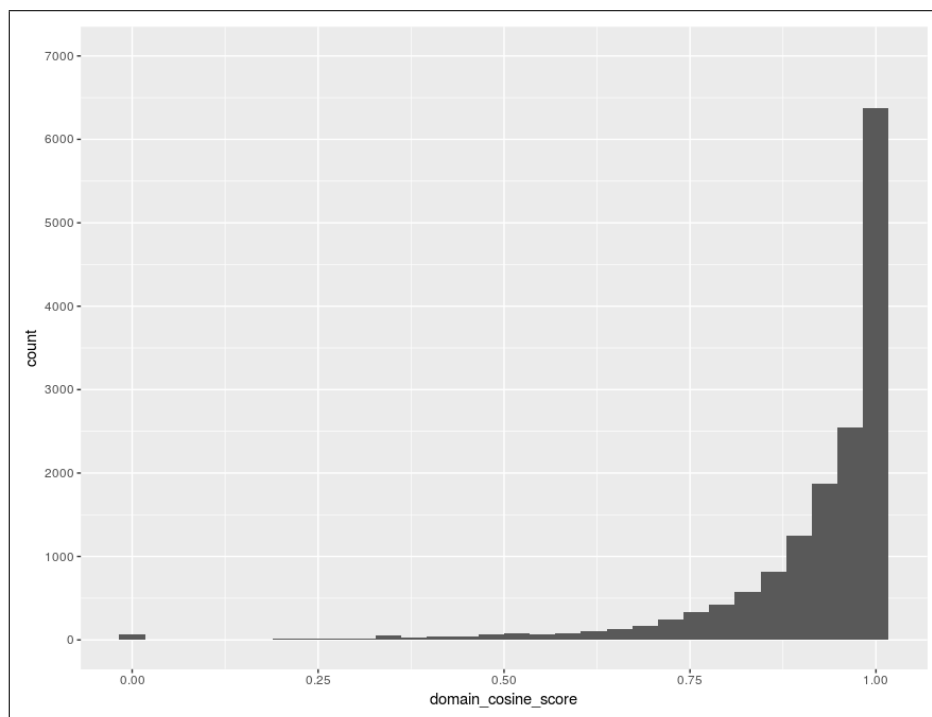
18

Figure 13: Distribution of domain composition based cosine scores for the lgf5 families
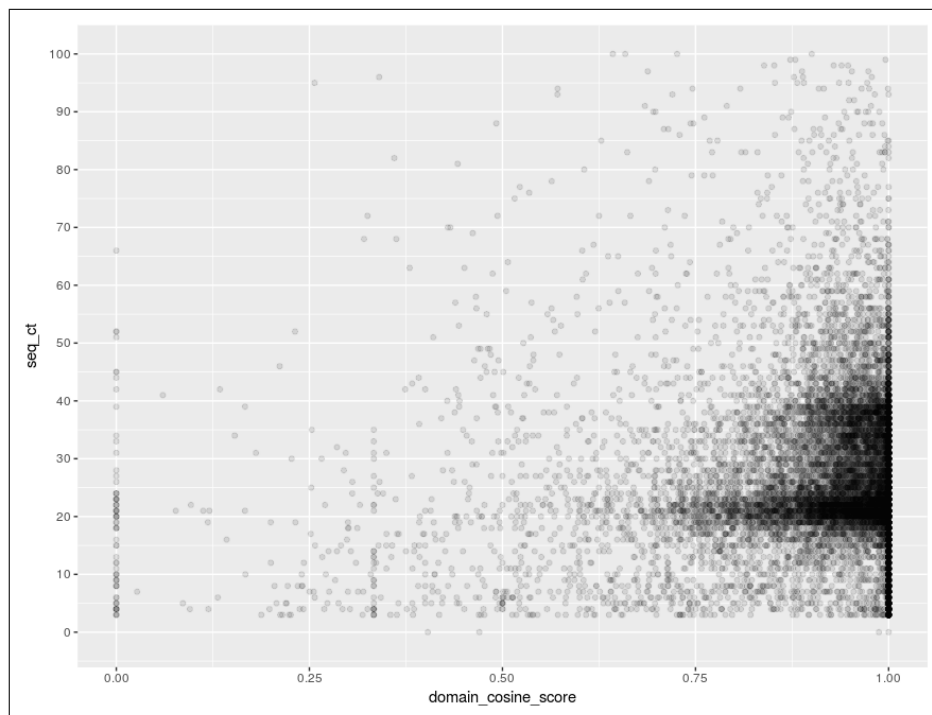
Figure 14: Plot of domain composition based cosine scores v/s ingroup sequence counts for lgf5 families

The scatter-plot (figure: 14) showing the relationship between the cosine score and family size was also obtained to study the relationship between cosine score and number of ingroup sequences in the family. Here too, the scores correlate well with family sizes, with higher scores for families with number of ingroup sequences between 15-45.

A scatter-plot (figure: 15) showing the relationship between the tree based precision scores and domain cosine scores was also obtained. Majority of families with high tree based precision scores also have high cosine scores. In addition, two types of families can be seen from the plot viz. families with high tree based precision scores and lower domain scores and vice versa. The former are those that seem to evolve rapidly through domain insertion/deletion. The latter are those that seem to evolve slowly conserving domain compositions over longer evolutionary periods which could be the reason for the over-clustering.

Both the domain based family scoring methods have been developed into a PYTHON tool that accepts directory containing family fasta files and the Pfam database, to assign domain composition scores to families.
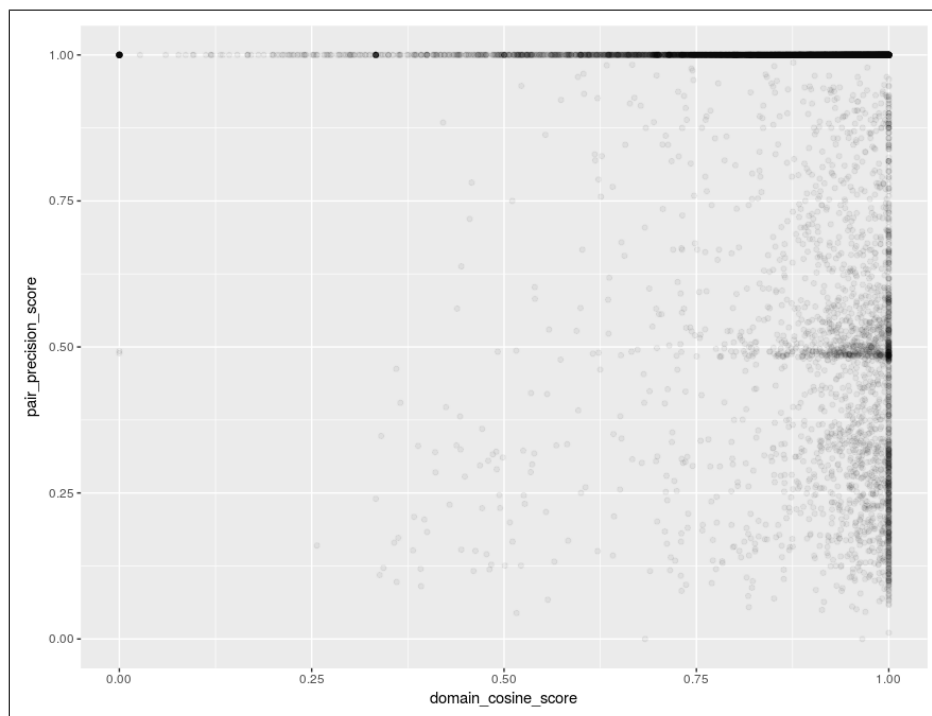
Figure 15: Plot of domain composition based cosine score v/s tree based precision scores for lgf5 families

5.2.3 Significance

Here we explore different domain based family scores mainly to detect over-clustering in families. Assuming that the species under study are closely related, the domain compositions for all the sequences in the family are expected to be largely conserved. The domain feature vector based cosine score seems to correlate better with family size and the tree based precision score as compared to the jaccard score.

The domain based scoring method can be used to detect candidate over-clustered families in conjunction with the tree based precision scores. This scoring method can also help in detecting rapidly evolving families.

5.2.4 Future work

This scoring method will be used in the conjunction with the tree based precision score for detecting and improving legume over-clustered families and for characterizing families. For example, large families with low tree-based precision scores and low domain composition scores are prime candidates for over-clustered families. Also, large families with low tree-based precision scores and high domain composition scores are slowly evolving families that could be over-clustered due to their highly conserved domain architectures over longer evolutionary periods.

It will also be interesting to study biological characteristics of families with high tree-based precision scores and low domain composition scores as these are the rapidly evolving families that have gained or lost protein domains within short evolutionary periods.

6 Objective 3: Building high-quality consensus set of families by comparing family sets

Well-defined families can be detected consistently by multiple family building methods. This workflow involves comparing the 2 sets of families, obtained from 2 different family building approaches, to establish family correspondences between the 2 sets i.e. which family from one set best corresponds to which family from the second set. Degrees of overlaps between corresponding families, between the two sets, can be used as measure of family consistency.

This workflow is used to compare set of Ks-based legume families to a set of families obtained from OrthoFinder, to obtain a set of high-quality consensus families. OrthoFinder is one of the latest family building methods that is proven to beat popular family construction methods including OrthoMCL. The two sets are compared through their respective family HMM sets.

6.1 Methods

6.1.1 Building HMM sets for families

Gene families are built using the OrthoFinder algorithm from 11 legume proteomes with 2 outgroup proteomes. Similarly, legume families are also built using in-house Ks-based family building algorithm, from 14 legume proteomes with 4 outgroup proteomes. Two family HMM sets are obtained from both the sets of families using the hmmbuild program from the HMMER package.

6.1.2 Classifying common set of proteomes into families using HMM sets

A common set of 11 legumes proteomes with 2 outgroup proteomes are classified into both the family HMM sets using best matching family HMM strategy. Both HMM sets are searched against the common set of proteomes, using the hmmscan program from HMMER package. Each sequence is classified into any one family, from both the sets, for which the sequence aligns best (e-value $\leq$ 1e-5) with the corresponding family HMM.

6.1.3 Calculating family correspondence and family overlaps between the sets

For each family in both the sets, largest overlapping family in opposite set is obtained. If, for example, fam1 from set1 is the largest overlapping family for fam2 from set2 and vice-versa, then fam1 and fam2 are considered corresponding families from both the sets. The two overlap scores, fam1-fam2 overlap and fam2-fam1 overlap are also calculated where fam1-fam2 overlap score is the proportion of sequences in fam1 that overlap with fam2 and, similarly, fam2-fam1 overlap score is the proportion of sequences in fam2 that overlap with fam1. If both the scores are 1.0, for a pair of families, then the two families match exactly from the two sets.

6.2 Preliminary results

The Ks-based and OrthoFinder legume family sets were compared using the family comparison workflow to explore the family correspondence between the two sets.

Nine thousand one hundred sixty families were found to match exactly between both the sets (both overlap scores = 1.0). Eleven thousand five hundred fifty three families were found to match with more than 0.9 overlap score between both sets and 12934 families were found to match more that 0.8 overlap score.

Both the overlap scores for each corresponding family pair from both the sets were plotted against each other (figure: 16). The families near the bottom right of the plot, on the y-axis, are those where the Ks-based family lies entirely within the corresponding orthofinder family. Similarly, the families near top left of the plot, on the x-axis, are those families where the orthofinder family lies entirely within the corresponding Ks-based family.
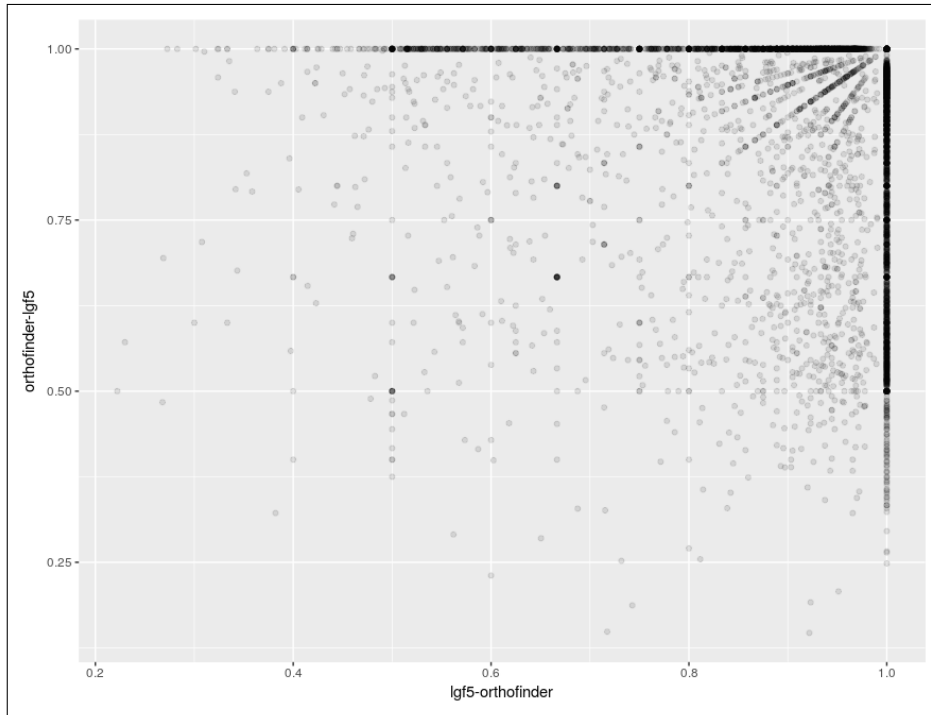

Figure 16: Plot of the two way overlap scores obtained from comparison between lgf5 and orthofinder family sets

Distribution of ingroup sequence counts (figure: 17) was also obtained from the families where both the overlap scores were ≥ 0.9, to study the family size distribution of families that are consistently detected by both family building

algorithms.
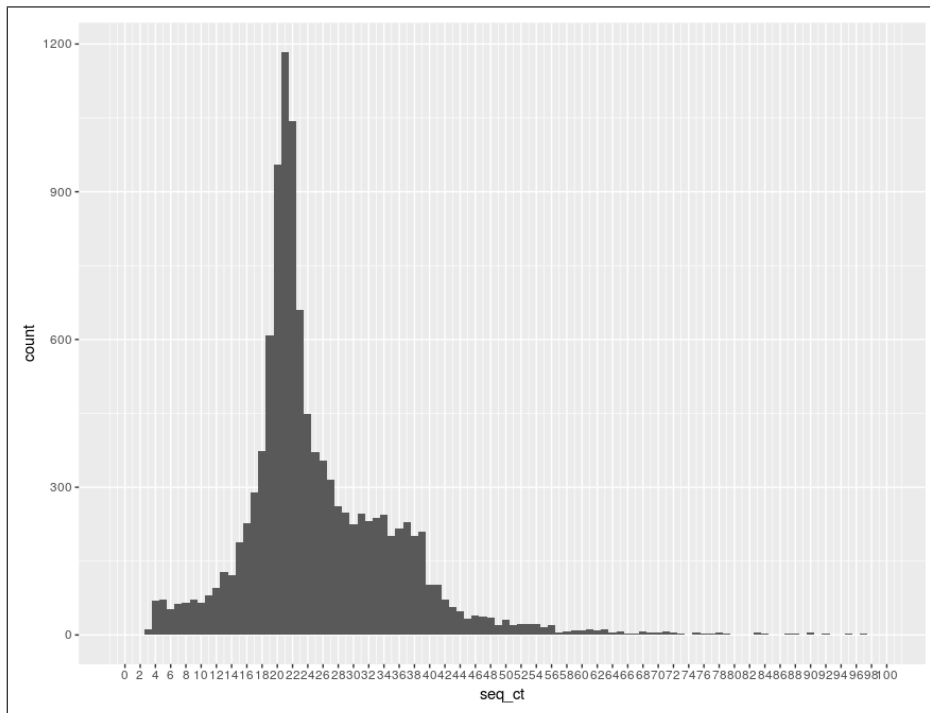
Figure 17: Distribution of ingroup sequence counts for lgf5 families with both overlaps scores ≥ 0.9

This workflow has been developed into a PYTHON tool which accepts 2 sets of HMMs corresponding to the 2 family sets and a single set of sequences to be classified into the two family sets.

6.3 Future work

The legume families that are found to overlap between the two sets can be released as high-quality set of families. Different sets can be assigned different confidence levels according to the overlaps found between the two sets.

7 Objective 4: Biological characterization of the legume families

This section involves exploring biological characteristics of different types of legume families. The legume ancestor is known to have undergone a WGD before divergence (Cannon et. al. 2014). However, some families seem to retain the duplicated gene and some families have lost the duplicate gene resulting from the WGD. As a result, legume families can be divided into distinct size classes - families with sizes between 10-30 and families between sizes 30-43. The former class of families are those that seem to have lost the duplicate gene and latter class have retained the duplicate gene. This work investigates if there is any biological basis to this phenomenon, using Gene Ontology (GO) enrichment analysis.

7.1 Methods

Legume families are divided into 4 groups corresponding to the different size classes (marked by vertical lines in fig.) with group 1 containing families with sizes between 0-10, group 2 containing families with sizes between 10-30, group 3 containing families with sizes between 30-43 and group 4 containing families with sizes 43 and above. Due to exceptionally large number of families in group 2, a subset of families in group 2 is selected for further processing. The distribution of family sizes (number of ingroup sequences) is shown in figure: 18 with the different groups marked with vertical lines. The dashed lines shows the subset of group 2 families.

Soybean (Glycine max) genes are extracted as representative genes from groups 1, 2-subset, 3 and 4, into separate lists. The gene lists are submitted for GO enrichment analysis on SoyBase (Grant et. al. 2009). Enriched ontology terms are obtained for all 4 genes lists corresponding to the 4 family groups.

7.2 Preliminary results

Distinct GO terms were found to be significantly enriched for all 4 groups of families. The enrichment for group 1 was less significant. This may be because group 1 may contain fragmented under-clustered families that could be merged into families from other 3 groups. Tables 1, 2, 3, 4 show the top 5 enriched GO terms for each of the size groups.

7.3 Future work

This analysis can be repeated on the high-confidence legume families obtained from the pair classification based family scoring workflow (merging under-clustered families), tree-based and domain-based scoring workflows (splitting over-clustered families) and the family sets comparison workflow (obtaining highly consistent families).
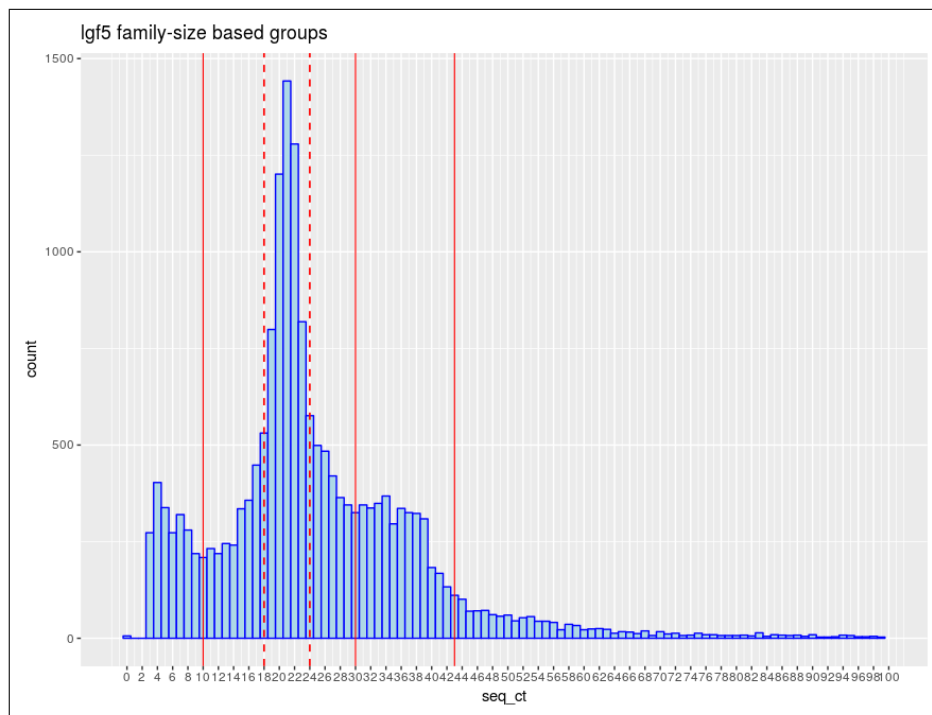
Figure 18: Distribution of ingroup sequence counts for lgf5 families

Applying domain based scoring on the high-confidence families can help in identifying slow-evolving/recent families (high jaccard and cosine domain composition scores) and fast-evolving/ancient families (low jaccard and cosine domain composition scores). GO enrichment analysis of these families can help in understanding the evolutionary pressures experienced by legume species during evolution.

GO id	P-value	GO description
GO:0009816	e-09	defense response to bacterium, incompatible interaction
GO:0010026	e-05	trichome differentiation
GO:0006355	e-04	regulation of transcription, DNA-dependent
GO:0015690	e-03	aluminum cation transport
GO:0009612	e-02	response to mechanical stimulus

Table 1: GO enrichment for Group 1 soybean genes

GO id	P-value	GO description
GO:0009658	e-45	chloroplast organization
GO:0010027	e-45	thylakoid membrane organization
GO:0019288	e-40	isopentenyl diphosphate biosynthetic process
GO:0006364	e-37	rRNA processing
GO:0019252	e-35	starch biosynthetic process

Table 2: GO enrichment for Group 2 subset soybean genes

GO id	P-value	GO description
GO:0006355	e-92	regulation of transcription, DNA-dependent
GO:0006468	e-25	protein phosphorylation
GO:0010200	e-21	response to chitin
GO:0010075	e-20	regulation of meristem growth
GO:0009651	e-19	response to salt stress

Table 3: GO enrichment for Group 3 soybean genes

GO id	P-value	GO description
GO:0006952	e-191	defense response
GO:0007165	e-66	signal transduction
GO:0006412	e-54	translation
GO:0055114	e-41	oxidation-reduction process
GO:0006865	e-40	amino acid transport

Table 4: GO enrichment for Group 4 soybean genes

References

- Fitch, W.M. 1970. Distinguishing homologous from analogous proteins. Syst. Zool. 19: 99113

- Fitch, W.M. 2000. Homology, a personal view on some of the problems. Trends Genet. 16: 227231

- Sonnhammer,E.L.L. and Koonin,E.V. (2002) Orthology, paralogy and proposed classification for paralog subtypes.Trends Genet.18, 619-620.

- Gardner, M.J., Hall, N., Fung, E., White, O., Berriman, M., Hyman, R.W., Carlton, J.M., Pain, A., Nelson, K.E., Bowman, S., et al. 2002. Genome sequence of the human malaria parasite Plasmodium falciparum. Nature 419: 498511.

- Kissinger, J.C., Brunk, B.P., Crabtree, J., Fraunholz, M.J., Gajria, B., Milgram, A.J., Pearson, D.S., Schug, J., Bahl, A., Diskin, S.J., et al. 2002. The Plasmodium genome database. Nature 419: 490492.

- Bahl, A., Brunk, B., Crabtree, J., Fraunholz, M.J., Gajria, B., Grant, G.R., Ginsburg, H., Gupta, D., Kissinger, J.C., Labo, P., et al. 2003. PlasmoDB: The Plasmodium Genome Resource. A database integrating experimental and computational data. Nucleic Acids Res. 31: 212215.

- Tatusov, R.L., Koonin, E.V., and Lipman, D.J. 1997. A genomic perspective on protein families. Science 278: 631637.

- Galperin, M.Y. and Koonin, E.V. 1999. Searching for drug targets in microbial genomes. Curr. Opin. Biotechnol. 10: 571578.

- Natale, D.A., Galperin, M.Y., Tatusov, R.L., and Koonin, E.V. 2000a. Using the COG database to improve gene recognition in complete genomes. Genetica 108: 917.

- Natale, D.A., Shankavaram, U.T., Galperin, M.Y., Wolf, Y.I., Aravind, L., and Koonin, E.V. 2000b. Towards understanding the first genome sequence of a crenarchaeon by genome annotation using clusters of orthologous groups of proteins (COGs). Genome Biol. 1: research0009.10009.19.

- Forterre, P. 2002. A hot story from comparative genomics: Reverse gyrase is the only hyperthermophile-specific protein. Trends Genet. 18: 236237.

- Tatusov, R.L., Galperin, M.Y., Natale, D.A., and Koonin, E.V. 2000. The COG database: A tool for genome-scale analysis of protein functions and evolution. Nucleic Acids Res. 28: 3336.

- Tatusov, R.L., Natale, D.A., Garkavtsev, I.V., Tatusova, T.A., Shankavaram, U.T., Rao, B.S., Kiryutin, B., Galperin, M.Y., Fedorova, N.D., and Koonin, E.V. 2001. The COG database: New developments in phylogenetic classification of proteins from complete genomes. Nucleic Acids Res. 29: 2228.

- Alexeyenko, A., Tamas, I., Liu, G. and Sonnhammer, E.L., 2006. Automatic clustering of orthologs and inparalogs shared by multiple proteomes. Bioinformatics, 22(14), pp.e9-e15.

- Remm, M., Storm, C.E. and Sonnhammer, E.L., 2001. Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. Journal of molecular biology, 314(5), pp.1041-1052.

- Roth, A.C., Gonnet, G.H. and Dessimoz, C., 2008. Algorithm of OMA for large-scale orthology inference. BMC bioinformatics, 9(1), p.518.

- Altenhoff AM, Schneider A, Gonnet GH, Dessimoz C. OMA 2011: orthology inference among 1000 complete genomes. Nucleic Acids Res. 2011;39:D28994.

- Li, L., Stoeckert, C.J. and Roos, D.S., 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome research, 13(9), pp.2178-2189.

- Emms, D.M. and Kelly, S., 2015. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome biology, 16(1), p.157.

- Lewis, G.P., 2005. Legumes of the World. Royal Botanic Gardens Kew.

- Blanc, G. and Wolfe, K.H., 2004. Functional divergence of duplicated genes formed by polyploidy during Arabidopsis evolution. The Plant Cell, 16(7), pp.1679-1691.

- Schlueter, J.A., Dixon, P., Granger, C., Grant, D., Clark, L., Doyle, J.J. and Shoemaker, R.C., 2004. Mining EST databases to resolve evolutionary events in major crop species. Genome, 47(5), pp.868-876.

- Pfeil, B.E., Schlueter, J.A., Shoemaker, R.C. and Doyle, J.J., 2005. Placing paleopolyploidy in relation to taxon divergence: a phylogenetic analysis in legumes using 39 gene families. Systematic Biology, 54(3), pp.441-454.

- Cannon, S.B., Sterck, L., Rombauts, S., Sato, S., Cheung, F., Gouzy, J., Wang, X., Mudge, J., Vasdewani, J., Schiex, T. and Spannagl, M., 2006. Legume genome evolution viewed through the Medicago truncatula and Lotus japonicus genomes. Proceedings of the National Academy of Sciences, 103(40), pp.14959-14964.

- Bertioli, D.J., Moretzsohn, M.C., Madsen, L.H., Sandal, N., Leal-Bertioli, S.C., Guimares, P.M., Hougaard, B.K., Fredslund, J., Schauser, L., Nielsen, A.M. and Sato, S., 2009. An analysis of synteny of Arachis with Lotus and Medicago sheds new light on the structure, stability and evolution of legume genomes. BMC genomics, 10(1), p.45.

- Cannon, S.B., McKain, M.R., Harkess, A., Nelson, M.N., Dash, S., Deyholos, M.K., Peng, Y., Joyce, B., Stewart Jr, C.N., Rolf, M. and Kutchan, T., 2014. Multiple polyploidy events in the early radiation of nodulating and nonnodulating legumes. Molecular biology and evolution, 32(1), pp.193-210.

- Kroc, M., Koczyk, G., wicicki, W., Kilian, A. and Nelson, M.N., 2014. New evidence of ancestral polyploidy in the Genistoid legume Lupinus angustifolius L.(narrow-leafed lupin). Theoretical and applied genetics, 127(5), pp.1237-1249.

- Eddy, S., 1992. HMMER users guide. Department of Genetics, Washington University School of Medicine, 2(1).

- Davis, J. and Goadrich, M., 2006, June. The relationship between Precision-Recall and ROC curves. In Proceedings of the 23rd international conference on Machine learning (pp. 233-240). ACM.

- Byrne, K.P. and Wolfe, K.H., 2005. The Yeast Gene Order Browser: combining curated homology and syntenic context reveals gene fate in polyploid species. Genome research, 15(10), pp.1456-1461.

- Li, H., Coghlan, A., Ruan, J., Coin, L.J., Heriche, J.K., Osmotherly, L., Li, R., Liu, T., Zhang, Z., Bolund, L. and Wong, G.K.S., 2006. TreeFam: a curated database of phylogenetic trees of animal gene families. Nucleic acids research, 34(suppl_1), pp.D572-D580

- Van der Heijden, R.T., Snel, B., Van Noort, V. and Huynen, M.A., 2007. Orthology prediction at scalable resolution by phylogenetic tree analysis. BMC bioinformatics, 8(1), p.83.

- Wapinski, I., Pfeffer, A., Friedman, N. and Regev, A., 2007. Automatic genome-wide reconstruction of phylogenetic gene trees. Bioinformatics, 23(13), pp.i549-i558.

- Hubbard, T.J., Aken, B.L., Beal, K., Ballester, B., Cccamo, M., Chen, Y., Clarke, L., Coates, G., Cunningham, F., Cutts, T. and Down, T., 2006. Ensembl 2007. Nucleic acids research, 35(suppl_1), pp.D610-D617.

- Dehal, P.S. and Boore, J.L., 2006. A phylogenomic gene cluster resource: the Phylogenetically Inferred Groups (PhIGs) database. BMC bioinformatics, 7(1), p.201.

- Finn, R.D., Tate, J., Mistry, J., Coggill, P.C., Sammut, S.J., Hotz, H.R., Ceric, G., Forslund, K., Eddy, S.R., Sonnhammer, E.L. and Bateman, A., 2007. The Pfam protein families database. Nucleic acids research, 36(suppl_1), pp.D281-D288.

- Falquet, L., Pagni, M., Bucher, P., Hulo, N., Sigrist, C.J., Hofmann, K. and Bairoch, A., 2002. The PROSITE database, its status in 2002. Nucleic acids research, 30(1), pp.235-238. Letunic, I., Doerks, T. and Bork, P., 2011. SMART 7: recent updates to the protein domain annotation resource. Nucleic acids research, 40(D1), pp.D302-D305.

- Marchler-Bauer, A., Lu, S., Anderson, J.B., Chitsaz, F., Derbyshire, M.K., DeWeese-Scott, C., Fong, J.H., Geer, L.Y., Geer, R.C., Gonzales, N.R. and Gwadz, M., 2010. CDD: a Conserved Domain Database for the functional annotation of proteins. Nucleic acids research, 39(suppl_1), pp.D225-D229.

- Song, N., Sedgewick, R.D. and Durand, D., 2007. Domain architecture comparison for multidomain homology identification. Journal of Computational Biology, 14(4), pp.496-516.

- Bitard-Feildel, T., Kemena, C., Greenwood, J.M. and Bornberg-Bauer, E., 2015. Domain similarity based orthology detection. BMC bioinformatics, 16(1), p.154.

- Mistry, J., Bateman, A. and Finn, R.D., 2007. Predicting active site residue annotations in the Pfam database. BMC bioinformatics, 8(1), p.298.

- Grant, D., Nelson, R.T., Cannon, S.B. and Shoemaker, R.C., 2009. SoyBase, the USDA-ARS soybean genetics and genomics database. Nucleic acids research, 38(suppl_1), pp.D843-D846.