

# **3PFDB – Database of best representative PSSM Profiles of Protein Families**

**K. Shameer, P. Nagarajan, K. Gaurav and R. Sowdhamini\***

**<sup>1</sup>National Centre for Biological Sciences  
Tata Institute of Fundamental Research  
GKVK Campus  
Bellary Road  
Bangalore 560 065  
INDIA**

**\* Author for correspondence; email: [mini@ncbs.res.in](mailto:mini@ncbs.res.in); Phone: +91-80-23666250; FAX:  
+91-80-23636462**

# **Abstract**

## **Background**

Protein families could be related to each other at broad levels that group them as superfamilies. These relationships are harder to detect at the sequence level due to high evolutionary divergence. Sequence searches are strongly directed and influenced by the best representatives of families that are viewed as starting points. PSSMs are useful approximations and mathematical representations of protein alignments, with wide array of applications in bioinformatics approaches like remote homology detection, protein family analysis, detection of new members and evolutionary modelling. Computationally intensive searches have been performed using the neural network based sensitive sequence search method called FASSM to identify the best representative PSSMs for families reported in Pfam database version 22.

## **Results**

A database of protein family-specific best representative PSSM profiles called 3PFDB has been developed. PSSM profiles in 3PFDB are curated using performance of individual sequence as a reference sequence in a rigorous scoring and coverage analysis approach using FASSM. We have assessed the suitability of 10, 85,588 sequences derived from seed or full alignments from Pfam database (Version 22). Coverage analysis using FASSM method is used as the filtering step to identify the best representative sequence, starting from full length or domain sequences to generate the final profile for a given family. 3PFDB is a collection of best representative PSSM profiles of 8,524 protein families from Pfam database.

## **Conclusion**

Availability of such a curated database of PSI-BLAST derived PSSMs for 91.4% of current Pfam family will be a useful resource for the community to perform detailed and specific analysis using family-specific, best-representative PSSM profiles in 3PFDB. 3PFDB can be accessed using the URL: <http://caps.ncbs.res.in/3pfdb>

## Background

Sensitive sequence search techniques play a vital role in enhanced function annotation approaches for several gene products in the post genomic era. The deluge of sequence data generated by high-throughput experiments need to be rapidly and effectively annotated using sensitive sequence search methods to understand the biological implications of individual sequences. Due to the practical inability of biochemical validation of large number of individual sequences from genome projects, bioinformatics tools are extensively developed and applied to enhance the function annotation of sequence and structural data [1-5]. BLAST [5] suite of programs are the first choice for such annotation of individual protein sequences based on homology and sequence conservation parameters. Position Specific Iterative BLAST (PSI- BLAST) [5] is one of the best flavours among the BLAST programs that offer a sensitive sequence search method for searching for homologous sequences and representing the amino acid conservation at different alignment positions into mathematical patterns using Position Specific Scoring Matrices (PSSM).

PSSM [6-8] is an useful approximation of sequence alignments that can be easily integrated in to a variety of tools and can be easily included in custom-code software [9, 10]. PSI-BLAST-generated position-specific scoring matrices can be used in a wide variety of application domains in bioinformatics like pattern recognition, machine learning, database searches, remote homology detection, prediction of transcription factors etc. In this paper, we report the availability of a database of best-representative PSSM profiles built on Pfam alignments subsequent to extensive analysis of individual members in a sequence family using FASSM (Function Association using Sequence & Structure Motifs) method [9].

In an earlier study, we have shown that FASSM method [9] can be used for validation by rigorous benchmarking studies. FASSM examines the sequence conservation and positions of protein family signatures or motifs for the annotation of protein sequences and to facilitate the analysis of their domains. Residues that characterize motifs at different alignment positions can also be identified using PSIMOT option in FASSM algorithm. FASSM method is driven by a neural network routine and was shown to be useful for difficult relationships such as discontinuous domains during whole-genome surveys and is demonstrated to perform accurate family associations at sequence identities as low as 15%. In the present instance,

FASSM algorithm and coverage analysis based on FASSM scoring is used to assess the ability of a sequence in a given protein family to generate the best-representative PSSM profiles. A database of “Best Representative PSSM profiles” (BRPs) of protein families (3PFDB) [11] is developed using a computationally intensive data-curation protocol that assessed 1.08 million PSI-BLAST generated PSSMs to identify the BRPs for 8,524 Pfam families. We also propose strategies for dealing with Pfam families where the associations of BRPs were not straightforward.

## Construction and content

### Data Curation:

Family specific best representative PSSM profiles in 3PFDB are identified using a computationally intensive exploratory search protocol. Every sequence in the (seed or full) alignment of a given protein family is given a chance to be the reference sequence and coverage analysis is performed using individual FASSM runs. Simplified graphical representation of the approach used to curate BRP of Pfam family PF00001 is provided in Figure 1. Different approaches based on seed and full datasets are followed to assess the suitability of a profile to be included in 3PFDB as the best-representative of a given family. . We have used Pfam version 22 (October 2007) for the data curation and 3PFDB database development. In this search protocol, we have successfully identified BRPs of 91.4 % of Pfam families in release Pfam 22. Detailed flow chart of the data curation steps is provided in Figure 2.

Steps in 3PFDB data curation:

1. For a given Pfam family, take all seed sequence dataset and generate PSSM profile using PSI-BLAST search against the seed alignment.
2. Upload these profiles to FASSM for assessment.
3. Use individual sequence from independent dataset and search using FASSM method.
4. Repeat the searches for all members in independent dataset (independent dataset was generated by removing seed sequence from the full alignment).
5. Each seed sequence was considered as a query (reference) sequence and searched against the sequence database generated from all of the seed sequences using PSI-BLAST to generate PSSM profiles.
6. These profiles were then uploaded to FASSM profile library.

7. Each of the full-length independent sequence was given as a query to FASSM to annotate the query to a particular seed sequence with a probability score.
8. The PSSM profile with the highest probability score to independent sequence was considered as the representative for the independent sequence.
9. PSSM profiles for all independent sequences were collected, PSSM profile with which represents independent sequence for more number of times will be considered as the BRP for a particular family.
10. For each family BRP are uploaded to the database.

### **Coverage analysis:**

Coverage of a particular PSSM profile was calculated from the ratio between the numbers of independent sequence it annotates to the total number of independent sequences in the family.

The coverage analysis formula is given in Figure 3.

Steps in Coverage analysis:

1. To retain BRP as representative for the family it should have a coverage value above the threshold limit of 50.
2. In some families BRP fails to cross the threshold limit, in those families the PSSM profile having highest coverage value was considered as the BRP.
3. In some families none of the PSSM profiles have coverage value above the threshold; in such cases, we have used only the domain of interest and generated independent sequences and re-examined the coverage.
4. In other instances, we have generated PSSM profiles from all the sequences available in the family (full set) and used all the sequence (domain length) as query for FASSM run.
5. Coverage was calculated for all the profiles, BRP was selected based on the highest coverage value.

### **Best-Representative PSSM profile (BRP) protein families:**

We introduce a new concept called BRP in 3PFDB. 3PFDB is developed as a result of an attempt to generate single PSSM profile for any given protein families. BRP of a given family is generated by the curation and coverage analysis method explained earlier. BRP is generated from the reference sequence that encapsulates all the important information of a diverse or highly similar family in to one single profile. BRP will be useful for researchers interested to perform large-scale protein family analysis. Protein family is a convenient level of sequence and structure based organization at which a group of proteins can be grouped to a

family based on different features like domain, sequence conservation, functional motifs, and structural similarity. Each member of a protein family will agree with similar features, still a protein family can have a wide-variety of members ranging from highly similar to highly diverse members. For example, PCA plot, hmm logo and alignment to describe the diversity of Regulators of G-protein Signalling family [12] (RGS family, Pfam ID: PF00615) is given in Figure 4. RGS proteins are multi-functional proteins with major role in signal transduction. The plot is generated using normalised alignment score from MALIGN [13] using GNUPLOT [14]. The alignment is curated from 'seed' alignment from Pfam and full length sequence is used for the generation of alignment. The plot clearly depicts the diversity within a protein family.

### **About 3PFDB database:**

In the current version of 3PFDB, 9318 Pfam families were analysed and best-representative profiles were identified for 8,524 families. The remaining 794 families were excluded from the database due to its poor performance in the data curation steps. On further analysis of this excluded dataset, we have observed that due to the large number of sequences in independent sequence, seed-based PSSM profile was not able to annotate all the sequences in the given family and the average family coverage have been fallen below 50%. As we set 50% as the cut-off for the family coverage, this family will not be included in the database. Another reason for the exclusion is that the individual PSSM profiles of the family are not having 50% coverage value. In this scenario, the profile of a given family is unable to annotate half of the family members. List of protein families available in the current version of 3PFDB [15] and list of excluded families [16] are provided for the easy access of the datasets. The current version of 3PFDB is corresponding to Pfam version 22 and 3PFDB will be updated periodically in response to the availability of newer versions of Pfam. A short delay in setting up the new version is anticipated due to the computationally intensive protocol used in the data curation steps. If users would like to perform BRP for custom generated alignments, users can contact the corresponding author for the FASSM program and other scripts used for the data curation.

### **Database Design:**

3PFDB is developed on a MySQL [17] backend. Server side CGI scripts are coded in Perl [18]. Web interface is developed using HTML, and JavaScript. FASSM scripts are coded using a combination of C++ and Perl. ANNIE [19] version 0.5 neural network package was used

to build neural network architecture. Blast version 2.2.16 [20] is used for PSI-BLAST [5] run and generation of PSSM profiles. BLAST generated alignments are converted in to PIR format using custom-Perl script. HMMER [21, 22] version 2.3.2 – is used to create the hmm models. The normalised alignment scores to generate PCA plots are obtained using ‘pca’ routine from MALIGN version 4.0 [13]. Normalised alignment scores are used to generate the plots using GNUPLOT 4.2 [14]. Database architecture of 3PFDB is provided in Figure 5.

### **Computing Time:**

The exploratory-search to identify best-representative PSSMs for 9318 Pfam families was performed on a 32-node cluster powered by Athlon 64-bit Quad-core processors running on a CentOS operating system version 4. 10, 85,588 rigorous PSI-BLAST searches were performed on the cluster to identify the best-representative PSSMs for 8,524 Pfam families. To perform the PSI-BLAST search, 5 months of CPU hours utilised to perform the primary data curation step in the development of 3PFDB. 8,524 ‘hmmbuild’ runs also were performed to generate hmm for the qualified family members.

### **Database Content:**

- FASSM based coverage analysis results
- PSIMOT-Motifs extracted using PSIMOT routine of FASSM
- PSIMOT Motifs marked on PSSM
- Sequence based PCA plot of the protein family
- Alignment of protein family in PIR format
- Download PSSM, HMM model and alignment
- Details about Pfam families
- Search and download options using Pfam family name, description and Pfam2GO annotations

### **Utility:**

3PFDB offers a unique collection of best-representative PSSM profiles. For each entry in the 3PFDB, following information is provided. FASSM based coverage analysis results, PSIMOT-Motifs extracted using PSIMOT routine of FASSM, PSIMOT-Motifs marked on PSSM, Sequence based PCA plot of the protein family, alignment of protein family based on best-representative and options are also available to family-specific best-representative

PSSM, HMM models and alignment in PIR format. 3PFDB also offers two text based search options to search and retrieve PSSM profiles using different set of key words. The searches are designed using Pfam description and Gene Ontology [23] based Pfam2GO[24] data. Pfam2GO is a useful way to map Pfam entries to GO. User can query 3PFDB using Pfam ID, Pfam description, Pfam short-description and Gene Ontology [23] related terms like GO ID or description. User can search the database using key words related to Pfam description and Pfam2GO annotation [24] and retrieve all the profiles that related to the key. As family specific and function-specific analysis is gaining importance in bioinformatics, availability of search engines to query 3PFDB using Pfam description and Gene Ontology will be useful.

## **Discussion and Conclusion:**

Several tools and databases employ PSSMs for different application in bioinformatics. These include but not limited to homology searches, pattern search, function assignment, function annotation, transcription factor binding site prediction, protein family classification using machine learning approaches like support vector machine. PSSMs are employed in several bioinformatics studies in different applications, for example predicting cyclin protein sequences [25], predictions of human, mouse and monkey MHC class I affinities for peptides [26], prediction method for virulent proteins in bacterial pathogens [27], sequence alignment and fold recognition with a custom scoring function [28], sequence-based prediction of DNA-binding residues in DNA-binding proteins [29], prediction of subcellular localization of gram-negative bacteria proteins [30]. Bioinformatics tools and databases like PROSITE [31], PRINT [32], BLOCKS [33] etc. employs PSSMs for pattern recognition based applications. CDD [34] provides a collection of PSSM profiles for Pfam families and MulPSSM [35] is another related resource that use multiple PSSMs corresponding to a given alignment and variable reference sequences. Databases like MULPSSM [39] have demonstrated the effectiveness of searching exhaustively, from different starting points and query sequences, in order to improve coverage. However, the total number of protein sequence domain families in databases like PFAM [36, 37] is far too high to handle all individual sequences. Owing to distant relationships and huge sequence dispersion within protein families, it is not always easy to find representative sequences. The concept of ‘seed’ sequences within Pfam databases is useful but does not, for many protein families, assure high and uniform coverage.

The Pfam database [36, 37] is a large collection of protein families, each represented by multiple sequence alignments and hidden Markov models (HMMs). Pfam is a database of sequence based protein domain families. Each family is represented by multiple sequence alignment and Hidden Markov Models. Pfam database is divided into two levels depending upon the quality of the families as Pfam-A and Pfam-B. Pfam-A is derived from the UniprotKB [38] derived sequence database 'Pfamseq'. Each Pfam-A family consists of a curated 'seed alignment' containing a small set of representative members of the family, profile hidden Markov models (profile HMMs) built from the seed alignment, and an automatically generated full alignment, which contains all detectable protein sequences belonging to the family, as defined by profile HMM searches of primary sequence databases. Pfam-B families are un-annotated and lower quality automated alignments generated automatically from the non-redundant clusters of ADDA [39].

In the current version of 3PFDB [11], we have used the seed alignment as the primary dataset to identify the BRP of a given protein family. Sequences from independent dataset (full dataset in a Pfam family without seed dataset) are used as the reference sequences to identify BRPs from 'seed dataset' for 7064 families. As this covers only 75% of Pfam version 20, we further used the sequences from 'full dataset' to identify BRPs of 1460 families (16%). As we assessed individual profile by its efficiency to annotate more than 50% of sequences in independent dataset in the case of 'seed dataset' and all sequences within a family in case of 'full dataset', the BRPs are one of the most authentic representation of a family in a profile format.

3PFDB provides a single best seed representative for the entire PFAM database and thereby immensely reduces the computational time for sequence searches and to establish relationships to the ever-expanding databases of sequence domain families. Further, the choice of the best seed representative using FASSM ensures best coverage since none of the seed sequences may uniformly attain high coverage. 3PFDB offers coverage analysis results for the Pfam family with other features of the database. For example, the coverage analysis results of the family [40] clearly indicates that the BRP using the sequence is derived from 'Q54LD1\_DICDI\_262\_386', this sequence was able to annotate 377 reference sequence with a coverage of 76.78%. Average coverage of this family starting from the seed sequences, however, was only 44.42%.

A new database of Best-Representative PSSM profiles (BRPs) of protein families called 3PFDB is developed. To the best of our knowledge, 3PFDB is first of its kind resource of BRPs generated using PSI-BLAST [5] and assessed through coverage analysis results of the sensitive sequence based annotation method FASSM [9]. PSSMs, alignments and HMM models available from 3PFDB can be extensively used for studies that require family-specific PSSM profiles.

#### **List of abbreviations used:**

3PFDB – Database of best-representative PSSM Profiles of Protein families, FASSM – Function Association using Sequence and Structural Motifs, HMM – Hidden Markov Model, PSI-BLAST – Position Specific Iterative – BLAST.

#### **AUTHORS' CONTRIBUTIONS:**

RS conceived of the study and discussed the approaches. PG had developed the FASSM algorithm; PN and KS had written the scripts and performed all the calculations; KS developed and organised the database. Both KS and PN wrote the first draft of the manuscript. KG and RS provided critical comments to the manuscript.

#### **ACKNOWLEDGEMENTS:**

We would like to thank Wellcome Trust (U.K.) and Department of Biotechnology (India) for financial support. We also thank NCBS (TIFR) for infrastructural and financial support.

#### **References:**

1. Whisstock JC, Lesk AM: **Prediction of protein function from protein sequence and structure.** *Q Rev Biophys* 2003, **36**(3):307-340.
2. Lee D, Redfern O, Orengo C: **Predicting protein function from sequence and structure.** *Nat Rev Mol Cell Biol* 2007, **8**(12):995-1005.
3. Laskowski RA, Thornton JM: **Understanding the molecular machinery of genetics through 3D structures.** *Nat Rev Genet* 2008, **9**(2):141-151.
4. Johnson MS, Srinivasan N, Sowdhamini R, Blundell TL: **Knowledge-based protein modeling.** *Crit Rev Biochem Mol Biol* 1994, **29**(1):1-68.
5. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25**(17):3389-3402.

6. Henikoff S: **Scores for sequence searches and alignments.** *Curr Opin Struct Biol* 1996, **6**(3):353-360.
7. Fogel GB: **Computational intelligence approaches for pattern discovery in biological systems.** *Brief Bioinform* 2008, **9**(4):307-316.
8. Gribskov M, McLachlan AD, Eisenberg D: **Profile analysis: detection of distantly related proteins.** *Proc Natl Acad Sci U S A* 1987, **84**(13):4355-4358.
9. Gaurav K, Gupta N, Sowdhamini R: **FASSM: enhanced function association in whole genome analysis using sequence and structural motifs.** *In Silico Biol* 2005, **5**(5-6):425-438.
10. Sandhya S, Chakrabarti S, Abhinandan KR, Sowdhamini R, Srinivasan N: **Assessment of a rigorous transitive profile based search method to detect remotely similar proteins.** *J Biomol Struct Dyn* 2005, **23**(3):283-298.
11. **3PFDB - Best representative PSSM Profiles of Protein Families**  
[<http://caps.ncbs.res.in/3pfdb>]
12. Watson N, Linder ME, Druey KM, Kehrl JH, Blumer KJ: **RGS family members: GTPase-activating proteins for heterotrimeric G-protein alpha-subunits.** *Nature* 1996, **383**(6596):172-175.
13. Johnson MS, Overington JP, Blundell TL: **Alignment and searching for common protein folds using a data bank of structural templates.** *J Mol Biol* 1993, **231**(3):735-752.
14. **GNUPLOT homepage** [<http://www.gnuplot.info/>]
15. **List of Pfam members with BRPs in 3PFDB (8, 524 families)**  
[<http://caps.ncbs.res.in/cgi-bin/mini/databases/3pfdb/browse.cgi?code=A>]
16. **List of Pfam members with out BRPs in 3PFDB (794 families)**  
[[http://caps.ncbs.res.in/cgi-bin/mini/databases/3pfdb/browse\\_mf.cgi?code=list](http://caps.ncbs.res.in/cgi-bin/mini/databases/3pfdb/browse_mf.cgi?code=list)]
17. **The MySQL Database** [<http://www.mysql.org>]
18. **Perl** [<http://www.perl.org>]
19. **ANNiE Artificial Neural Network Library** [<http://annie.sourceforge.net/>]
20. **BLAST version 2.2.16** [<ftp://ftp.ncbi.nlm.nih.gov/blast/executables/release/2.2.16/>]
21. Eddy SR: **Profile hidden Markov models.** *Bioinformatics* 1998, **14**(9):755-763.
22. **HMMER: biosequence analysis using profile hidden Markov models**  
[<http://hmmmer.janelia.org/>]
23. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT *et al*: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.** *Nat Genet* 2000, **25**(1):25-29.
24. **Pfam2GO** [<http://www.geneontology.org/external2go/pfam2go>]
25. Kalita MK, Nandal UK, Pattnaik A, Sivalingam A, Ramasamy G, Kumar M, Raghava GP, Gupta D: **CyclinPred: a SVM-based method for predicting cyclin protein sequences.** *PLoS ONE* 2008, **3**(7):e2605.
26. Lundegaard C, Lamberth K, Harndahl M, Buus S, Lund O, Nielsen M: **NetMHC-3.0: accurate web accessible predictions of human, mouse and monkey MHC class I affinities for peptides of length 8-11.** *Nucleic Acids Res* 2008, **36**(Web Server issue):W509-512.
27. Garg A, Gupta D: **VirulentPred: a SVM based prediction method for virulent proteins in bacterial pathogens.** *BMC Bioinformatics* 2008, **9**:62.
28. Dong E, Smith J, Heinze S, Alexander N, Meiler J: **BCL::Align-Sequence alignment and fold recognition with a custom scoring function online.** *Gene* 2008, **422**(1-2):41-46.

29. Hwang S, Gou Z, Kuznetsov IB: **DP-Bind: a web server for sequence-based prediction of DNA-binding residues in DNA-binding proteins.** *Bioinformatics* 2007, **23**(5):634-636.
30. Guo J, Lin Y, Liu X: **GNBSL: a new integrative system to predict the subcellular location for Gram-negative bacteria proteins.** *Proteomics* 2006, **6**(19):5099-5105.
31. Hulo N, Bairoch A, Bulliard V, Cerutti L, De Castro E, Langendijk-Genevaux PS, Pagni M, Sigrist CJ: **The PROSITE database.** *Nucleic Acids Res* 2006, **34**(Database issue):D227-230.
32. Attwood TK, Bradley P, Flower DR, Gaulton A, Maudling N, Mitchell AL, Moulton G, Nordle A, Paine K, Taylor P *et al*: **PRINTS and its automatic supplement, prePRINTS.** *Nucleic Acids Res* 2003, **31**(1):400-402.
33. Henikoff JG, Greene EA, Pietrokovski S, Henikoff S: **Increased coverage of protein families with the blocks database servers.** *Nucleic Acids Res* 2000, **28**(1):228-230.
34. Marchler-Bauer A, Anderson JB, Derbyshire MK, DeWeese-Scott C, Gonzales NR, Gwadz M, Hao L, He S, Hurwitz DI, Jackson JD *et al*: **CDD: a conserved domain database for interactive domain family analysis.** *Nucleic Acids Res* 2007, **35**(Database issue):D237-240.
35. Gowri VS, Krishnadev O, Swamy CS, Srinivasan N: **MulPSSM: a database of multiple position-specific scoring matrices of protein domain families.** *Nucleic Acids Res* 2006, **34**(Database issue):D243-246.
36. Finn RD, Tate J, Mistry J, Cogill PC, Sammut SJ, Hotz HR, Ceric G, Forslund K, Eddy SR, Sonnhammer EL *et al*: **The Pfam protein families database.** *Nucleic Acids Res* 2008, **36**(Database issue):D281-288.
37. Sammut SJ, Finn RD, Bateman A: **Pfam 10 years on: 10,000 families and still growing.** *Brief Bioinform* 2008, **9**(3):210-219.
38. **The universal protein resource (UniProt).** *Nucleic Acids Res* 2008, **36**(Database issue):D190-195.
39. Heger A, Holm L: **Exhaustive enumeration of protein domain families.** *J Mol Biol* 2003, **328**(3):749-767.
40. **Coverage analysis results for PF00615 in 3PFDB** [[http://caps.ncbs.res.in/cgi-bin/mini/databases/3pfdb/get\\_entry.cgi?id=PF00615](http://caps.ncbs.res.in/cgi-bin/mini/databases/3pfdb/get_entry.cgi?id=PF00615)]

## Figure Legends

Figure 1: Simplified graphical representation of the data curation to identify BRP of Pfam family PF00001

Figure 2: Detailed flow chart of the data curation steps in 3PFDB

Figure 3: 3PFDB – Coverage analysis formula

Figure 4: PCA plot, HMM logo and alignment of RGS family (PF00615) from 3PFDB

Figure 5: 3PFDB – Database architecture

# 3PFDB – PSSM Generation

PF00001  
SEED : 64  
FULL : 14615  
IND : 14551

- 64 PSSM profiles generated
- Independent sequences (14551) were used as query to search in FASSM with library of 64 profiles
- FASSM Score based coverage analysis

Pfam ID	Best Profile	Coverage of Best Profile	Average Coverage	Best Profile Based on FASSM Score	Coverage
PF00001	PE2R4_HUMAN_34_329	80.49	63.91	OPS3_DROME_75_338	82.97

Best Profile based on Reference  
PE2R4\_HUMAN\_34\_329 is added  
to 3PFDB

```
LAIAMNYLLT-----  
LAIAMNYLLT-----  
LQAAAYFLOVP-PEI  
LKAAGATL--DAPL  
VELARISLN-----
```

**3PFDB**

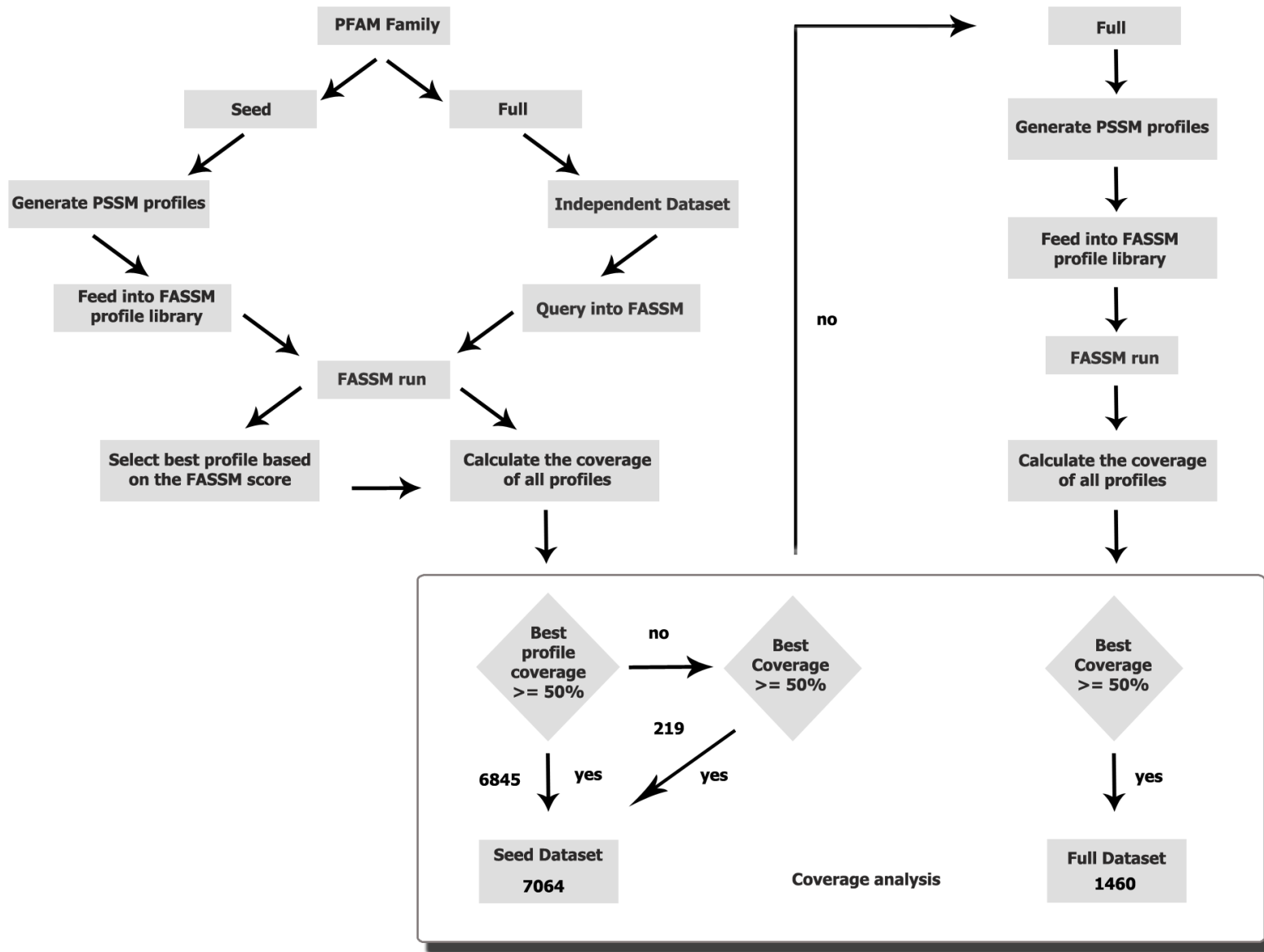


Figure 2

$$\text{Coverage of PSSM profile X of family Y} = \frac{\text{Total number of independent sequences annotated by PSSM profile X}}{\text{Total number of independent sequences in the family Y}}$$

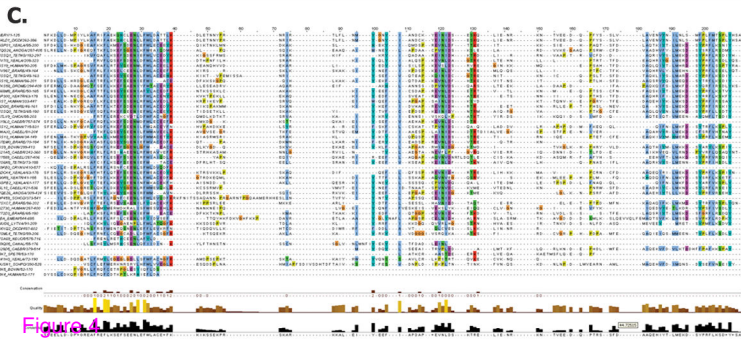
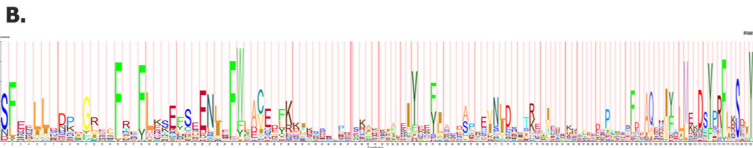
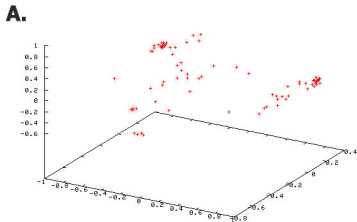


Figure 4

PSSM

Alignment

PSIMOT  
MotifsHMM  
Models

MySQL

Pfam Description

Coverage Analysis Results

PCA Images

Flat files

CGI Programs for search &amp; retrieve

Server Side Script

3PFDB - PSSM Profiles of Protein Families

Home Search Download Help Email

3PFDB - PSSM Profiles of Protein Families : Database Entry - PF00028

**Details:**  
 Pfam ID : PF00028  
 Pfam Domain Description : **Cadherin domain**  
 Curated from SEED dataset  
 Curated using FULL Sequence


**PSSM Based Coverage Results:**

Pfam ID	Best Profile	Coverage of Best Profile	Average Coverage	Best Profile Based on Coverage	Coverage
PF00028	FAT_DROME_392_485	89.71	43.99	FAT_DROME_392_485	89.71

**PSIMOT-Motifs Extracted using PSIMOT:**  
 There are 2 Motifs  
 Motif1 Start=5 Length=27  
 Motif2 Start=62 Length=8

**PSIMOT Motifs Marked on PSSM:**  
[Show PSSM](#)

**Sequence based PCA Plot of the family - PF00028 :**  
[Show PSSM](#)



**Alignment of Protein Family - PF00028 :**  
[Show Alignment](#)

**Download PSSM/HMM Models/Alignment**

PSSM	HMM	Alignment
<a href="#">Download</a>	<a href="#">Download</a>	<a href="#">Download</a>

HTML, JavaScript  
Dynamic Web Interface