

## Reviewer's report

**Title:** 3PFDB - Database of best representative PSSM Profiles of Protein Families

**Version:** 1 **Date:** 24 February 2009

**Reviewer:** Oruganty Krishnadev

### Reviewer's report:

The authors of this paper describe the creation of a database of 'best representative profiles' (BRPs) for Pfam families using PSI-BLAST for profile generation. The choice of representative sequences has till now been done in an ad hoc manner and hence, their work addresses an important question in the field. However, I am unable to convince myself that this is the most appropriate way to select BRPs

#### \*\*\*\*\* Major Compulsory Revisions \*\*\*\*\*

The definition of a 'best representative profile' (BRP) for a family is subjective and is dependent on the application which uses the profiles (that is, there is no data supporting the notion of a universally best representative for a family). In this paper, the authors define BRP as the profile with greatest coverage of 'independent database' (Pfam full dataset – Pfam seed dataset) using the program FASSM. The sensitivity of the underlying method thus plays a crucial role in the BRP generation (for example, HMMER would give 100% coverage for each of the families in Pfam). So, the points with which I am concerned are as follows:

1) Given that coverage of a family is the parameter used for selection of FASSM, is it established that FASSM is as sensitive as /more sensitive than RPS-BLAST (the analogous sequence-profile matching program in BLAST suite) ? The FASSM paper reports that it is 93% sensitive in comparison to IMPALA. How does it fare in comparison to RPS-BLAST and other commonly used programs ? This is important to consider since FASSM based analysis has failed to provide BRPs for a substantial number of families.

2) If, FASSM is not as sensitive as RPS-BLAST, then an analysis should be done using a random subset of Pfam families (maybe 10% of Pfam families) using RPS-BLAST. The BRPs selected using both methods (RPS-BLAST and FASSM) can then be assessed for 'family coverage' by querying the 'independent dataset' against both the databases. This will give an indication of the efficiency of BRPs selected using FASSM in comparison to BRPs selected using BLAST set of programs (necessary, since BLAST suite is more commonly used for remote homology searches, and would give the user a handle on what to expect when using BRPs present in 3PFDB).

NOTE : A preliminary analysis for a few families (7 to be precise) done at the reviewer's lab reveals that RPS-BLAST has greater/equal sensitivity than FASSM

in all cases, and the profiles selected are different than those given in 3PFDB. Also, in two cases (PF00485, PF09198), 3PFDB fails to find a BRP, whereas RPS-BLAST finds a BRP at 65% and 100% coverage respectively. In another case (PF00915), both RPS-BLAST and 3PFDB fail to provide a BRP.

3) If the above step indicates that RPS-BLAST would be a better choice for BRP generation, then the current work can still be considered important if it is shown that the BRPs generated using FASSM give better performance in other applications using PSSMs (for example in secondary structure prediction or in profile-profile matching procedures). This is the motivation with which the authors created the database.

\*\*\*\*\* Minor Essential Revisions \*\*\*\*\*

1) Figure 1 needs to be explained in more detail, either in the main text or in the figure legend. The figure reports two coverage values, in the third column "Coverage of Best profile" and the last column "Coverage". In this case, although the last column reports a higher coverage, the profile chosen is not what is given as "Best profile based on FASSM score". The discrepancy needs to be explained and the two coverage values should be defined unambiguously and the choice of BRP explained.

2) "Construction and Content" section, "Data curation" subsection should be revamped and the various steps should be defined unambiguously. Considering that this is the part where the analysis is explained in detail, it has to be clear and unambiguous. The various statements which are not very clear are :

2a) Steps 1-4 and Steps 5-7 seem redundant in the current version. Perhaps, they need to be explained in greater detail ?

2b) Step 8 suggests that only one PSSM, among the many family PSSMs, is chosen for each independent sequence. Is this true ? If not, consider revising the statement. If it is true, then how is average family coverage greater than 50% in families with greater than 2 seed sequences ?

2c) Step 9 is not very clear. Are the authors trying to say that the "Matching PSSM profiles for all independent sequences are collected" in the first part ? The second part is also not clear. Are the authors trying to say that "the PSSM profile which represents most independent sequences is considered as BRP" ?

3) "Coverage analysis" subsection, the steps are again not explained very clearly. The points which are not clear are :

3a) Step 1 and Step 2 can be combined to say that the PSSM which has the highest coverage is retained as the BRP. Obviously, this is not what the flowchart in Figure 2 suggests. Consider revising the statements. Also, in Step 2 perhaps, the statement should be "In some families, 'PSSMs' fail to cross the threshold limit, in those....." (PSSM used instead of BRP) ?

3b) Step 3 is not clear. Pfam families are based on an ad hoc functional domain definition. What do the authors mean by "using only domain of interest", and "generating independent sequences" ? The two terms have to be explained in greater detail.

3c) Two coverage values are indicated in the flowchart in Figure 2. In the main section, the word coverage is used for both 'coverage of profile' and 'family coverage'. The section would be clearer to understand if a clear distinction is made on which of the coverage values are being mentioned (if coverage on individual profile was taken into consideration).

3d) a) A small section or subsection on FASSM probability scores and the method used to find homologous PSSMs for a query sequence can be included, giving details about the 'coverage of profile' calculation.

4) Typographical or minor mistakes

4a) Background section, para 1, line 10. "(PSI-BLAST) is one of the best flavours among ...." Consider replacing 'flavours' with 'variant' or similar word.

4b) Discussion and conclusion section, para 2. first and third lines are redundant.

4c) Discussion and conclusion section, para 3, line 2. '...are used as the reference sequences to identify ...'. According to earlier pages, the independent sequence dataset sequences should not be reference sequences.

4d) Discussion and conclusion section, para 3, line 3. '.... this covers only 75% of Pfam version 20...'. Isn't it Pfam version 22 ?

\*\*\*\*\* Discretionary Revisions \*\*\*\*\*

PSSM generation has been done for most families using the seed sequences as representative sequence and querying against the seed database. It is not very clear to me if the sensitivity of the PSSMs is compromised because of the limited number of sequences available for most Pfam families (which can have very high sequence identity in some cases). Perhaps the authors can generate PSSMs using the seed sequence as the representative sequence and querying in the full dataset to obtain PSSMs for a family. It is not guaranteed to give better results and would take a long time, so perhaps it could be done on a smaller scale to see if it makes a difference to the family coverage values.

**Level of interest:** An article of importance in its field

**Quality of written English:** Needs some language corrections before being published

**Statistical review:** No, the manuscript does not need to be seen by a statistician.

**Declaration of competing interests:**

I declare that I have no competing interests.