Manual curation improves accuracy of Pfam families with amino acid compositional bias

Jaina Mistry[1], Robert D. Finn[2], Sean R. Eddy[2], Alex Bateman[1] and Marco Punta[1]

1. Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SA, UK.
2. HHMI Janelia Farm Research Campus, 19700 Helix Drive, Ashburn, VA 20147, USA.

1. Introduction

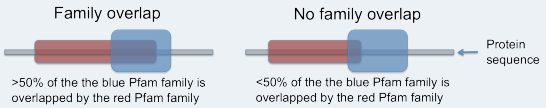
- Homology is widely used as a basis for transferring structural and functional annotation.
- Pfam is a database of profile hidden Markov model (HMM)-based families of homologous protein regions that are built and searched using the HMMER3 software [1].
- A significance threshold (also known as the gathering or GA threshold) for homology is manually selected for each family.
- Proteins are usually comprised of one or more discrete domains; as a quality control measure, we do not allow Pfam families to overlap.

2. Objective

- We tried to find out how much benefit we gain from manually curating family-specific significance thresholds. Selecting a fixed E-value threshold for all families would save time but might lead to lower specificity and/or sensitivity.

3. Method

- Family overlaps are defined as overlaps between two families that are not in the same Pfam clan, where the overlap covers at least 50% of one of the two matches involved.



- Pfam 26.0 HMMs were searched against the UniProt Knowledgebase (UniProtKB). Membership of each family was determined using different fixed bit score thresholds of 23, 25, 27 and 30, fixed E-value thresholds of 0.1, 0.05, 0.01 and 0.001, as well as Pfam GAs and Pfam GAs with a cap on the minimum E-value (matches with E-value >0.1 are ignored).
- Sequence coverage is the proportion of UniProtKB sequences that have a match to a Pfam family.
- When looking at regions with a biased amino acid composition in overlapping families (Figure 3), we used a greedy algorithm for assigning each overlap to a single family.
- For the seed alignments of all Pfam families, coiled-coil regions were predicted using ncoils [3], disordered regions using IUPred [4] and trans-membrane regions using Phobius [5].

4. Method validation

Validation of overlaps as proxies for false positives

- We use overlaps between Pfam families as proxies for false positives. Overlaps, however, could indicate genuine, previously unrecognised evolutionary links between families. We tested our hypothesis using the subset of overlap data where both families had a 3D structure.
- Depending on the fixed threshold, between 45 and 61% of overlaps that map to SCOP domains are found to be evolutionarily unrelated (**Figure 1**).
  - We will use the number of overlaps between Pfam families as a crude estimate of the number of false positives generated by each significance threshold.

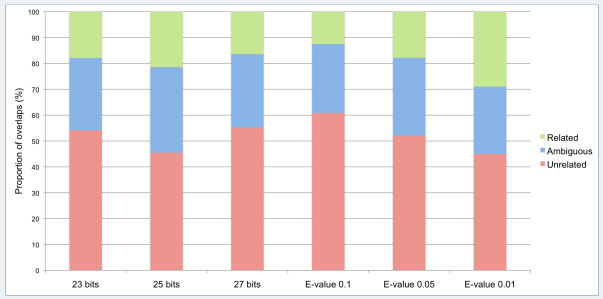


Figure 1. Predicted homology between pairs of families featuring an overlap, when both families have a structure in SCOP. Data shown for different fixed bit score and E-value thresholds. To determine which structures were evolutionarily related, evolutionarily unrelated, or ambiguous, we used a benchmarking system developed by Madera and Gough [5], based on the SCOP database.

5.1 Results

UniProtKB coverage and number of overlaps

- The Pfam GAs give a low estimated false positive rate compared to fixed bit score and E-value thresholds that give a similar UniProtKB coverage (**Figure 2**)

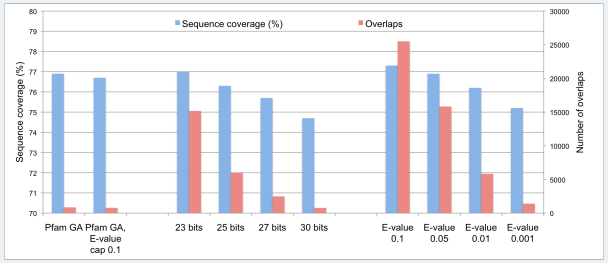


Figure 2. Overlaps and sequence coverage for Pfam GA and different fixed bit score and E-value thresholds.

5.2. Results

Investigation of families that overlap at a fixed E-value threshold of 0.05

- 377 families (2.8% of all Pfam families) contribute 90% of all overlaps.
- Families that contribute overlaps are enriched in trans-membrane and coiled-coil regions (**Figure 3**)

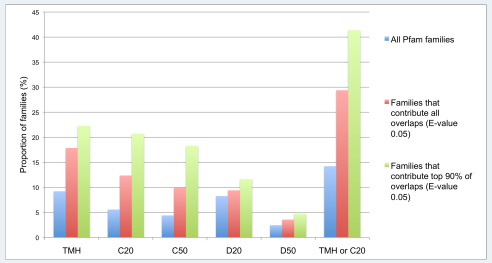


Figure 3. Proportion of families that have >2 transmembrane helices (TMH), consecutive coiled-coil regions of 20 (C20) and 50 (C50) residues, and consecutive disordered regions of 20 (D20) and 50 (D50) residues, in >50% of the members of their seed alignments.

6. Conclusions

- For most families, a fixed threshold is effective in defining homology, however in a small but significant number of Pfam families (3-5% of the total), manual curation seems able to greatly reduce the number of false positives.
- Several of the families in which there is a high estimated rate of false positives in fixed thresholds are enriched in low complexity regions such as trans-membrane and coiled-coil segments.

7. References

1. http://hmmer.janelia.org
2. http://www.russelllab.org/cgi-bin/coils/coils-svr.pl
3. http://iupred.enzim.hu
4. http://phobius.sbc.su.se
5. Madera, M. and Gough, J. (2002) A comparison of profile hidden Markov model procedures for remote homology detection. Nucleic Acids Res, 30:4321-4328.

Get a pdf of this poster