# Crowdsourcing Protein Family Database Curation

Matt Jeffryes[1], Maria Liakata[2], Alex Bateman[1]
[1] European Bioinformatics Institute (EMBL-EBI), Cambridge, United Kingdom
{mjj, agb}@ebi.ac.uk
[2] University of Warwick, Coventry, United Kingdom
m.liakata@warwick.ac.uk

*Abstract*— **We propose a novel method for crowdsourcing a protein family database. We discuss how we intend to identify novel groupings of proteins from user sequence similarity search, and how text mining will be applied to assist in annotation of these novel groupings, and more broadly as an enrichment of protein sequence similarity search results. We intend to use entity linking to identify literature which discusses proteins found in the search results, and present those publications which are likely to be the most useful to curators and sequence similarity search users alongside the sequence search results.**

***Keywords—crowdsourcing; biocuration; databases***

## I. Introduction

Protein families are groupings of proteins which have an evolutionary relationship. Pfam is a database of protein families, which has been maintained since 1996 [1]. During this time, Pfam has been a human curated database. Each protein family in the database is defined by an alignment which has been constructed by curation staff. Using this approach, Pfam has reached coverage of 47% of the residues in the protein sequence database upon which it is based [2]. Curators use the software package HMMER to construct Pfam families. For each family, a number of exemplar 'seed sequences' are aligned, HMMER is used to produce a sequence profile hidden Markov model (HMM). This HMM is then used to query *pfamseq*, the protein sequence database upon which Pfam is based. The regions in the database which are significant matches for the HMM are the members of the protein family [1].

HMMER is also a valuable tool for discovering families. It can perform sensitive sequence similarity searches against a target protein sequence database, and is able to find homologous proteins distantly related to the query sequence [3]. When used in this way, HMMER builds a sequence profile HMM *de novo* from the query sequence, and retrieves matching sequences from the sequence database. HMMER can either be run at the command line against a local database, or via the HMMER web service [4]. In addition to the sequence profile HMM which defines a family, Pfam curators annotate families with literature citations which give further evidence of the existence of the family, and cross-references to other databases, via Interpro.

Since 2011, Pfam has used Wikipedia as the primary repository for family annotation. If an article already exists, curators link it to the family, and the contents of the article is mirrored on the protein family's web page. For families which do not have a Wikipedia article, users are encouraged to create a new article [5]. This method of annotation was pioneered by Pfam's sister-database Rfam and has resulted in higher quality annotation. As an additional benefit, crowdsourcing annotation has increased the visibility of the databases to the scientific community [6].

Constructing a new Pfam entry requires that a curator searches the scientific literature for evidence of the existence of the new protein family, and for the possible function and structure of the members of the family. The curator's aim is to identify literature which mentions a member of the family, and the most useful papers will be those which go into detail about the protein's structure or function.

Identifying literature which mentions a particular protein is more complex than searching the literature for the protein's name. Biologists may use several different names and abbreviations for the same protein. Additionally, the same name may be shared across several species [7].

Previously, PubServer combined protein sequence similarity search using PSI-BLAST with literature retrieval [8]. Our approach differs in two ways. Firstly, PubServer retrieves only publications which are already attached to protein database entries, whereas our approach is to search the open-access subset of PubMed Central for relevant mentions of proteins. Secondly, our approach uses article full-text rather than being limited to title, abstract and MeSH terms.

Our aim is to introduce crowdsourcing to the construction of the families, and to facilitate easier annotation of families by the application of text mining.

## II. Proposed System

### A. Crowdsourcing Profile HMMs

HMMER can be used to identify new protein families. This could happen intentionally, when a Pfam curator makes additions to the database, but it can also occur incidentally, when a sequence similarity search user's query sequence happens to produce an HMM which matches against sequence regions which are as yet not matched by any existing Pfam family. It can also occur that a user's search matches a very high proportion of the sequences matched by an existing Pfam family, and further additional sequences which are likely to be homologous with the Pfam family. In this case, the HMM produced by the user's search is a potential improvement over the HMM which currently defines the family.

Presently, such potential improvements to Pfam will only be incorporated if the user identifies that they have found a grouping of proteins unknown to Pfam, and makes the effort to contact Pfam's helpdesk. We propose that users should be alerted to this situation and prompted to submit their search's sequence profile HMM to Pfam.

We believe that this has two advantages over an alternative system which would silently capture HMM's from sequence similarity search. Firstly, the user who identifies the new or improved family gets credit for their improvement, which we consider to be both an ethical necessity, and an opportunity for increasing community engagement. Secondly, we hypothesise that the user who performs a sequence similarity search which identifies a novel protein family is more likely to have the knowledge required to annotate that family with relevant literature or other metadata than Pfam curators are, resulting in higher quality annotation. This second advantage also applies over a potential alternative system which automatically performs sequence similarity search with random sequences which are not matched by any Pfam family. While this would likely produce many novel families, the curation workload to actually incorporate these families into Pfam would be high.

### B. *Crowdsourcing Annotation*

Once a user has been made aware that their sequence similarity search represents a potential improvement to Pfam, we would like to ease the process of adding annotations to their new family. In particular, we would like to highlight literature which is relevant to their sequence similarity search.

Presently, when researching a protein found by a sequence similarity search, users could look at the literature linked to the protein in a protein database such as UniProt. However, the literature citations for an entry are not intended to be exhaustive, but a representative sample [9]. It is also hard to use this method to search for literature which mentions multiple proteins found in a search. Moreover, relevant sections within known papers are not available. Our proposed system will use named entity recognition to identify literature mentioning protein sequences matched by a HMMER sequence similarity search result. We will then seek to extract the passages from the text which are most relevant for annotation and use these to rank the relevance of publications. We hypothesise that information about protein function, structure, homology, and phylogeny will be the most useful. Literature which mentions multiple proteins found by a single sequence similarity search is also likely relevant.

We anticipate that literature search will be useful not only to curators and users constructing new protein families, but also to sequence similarity search users in general, as a tool for researching homologues of an uncharacterised protein sequence.

Our proposed method is to first use the BANNER named entity recogniser, trained on the BioCreative II gene mention data set, to identify all possible mentions of genes and proteins within the open access subset of PubMed Central [10]. We will then train a ranking SVM classifier to identify the most relevant mentions for entities in the protein database UniProt. This method is based upon [11], but with a reversal in the direction of association: Instead of identifying candidate entities for each possible mention in the text, we will identify candidate mentions for each entity within the knowledge base.

We intend to train and evaluate this method by using the references attached to each entry in the manually annotated subset of UniProt, SwissProt. Each publication in the reference list of a SwissProt entry has a ‘scope’, which describes the elements of the entry that the publication has been cited for. For example, an entry may have one reference with a scope of ‘nucleotide sequence’, which describes the sequencing of the gene which codes for the protein. The proposed system can be evaluated by its ability to extract the elements of the manually curated reference list from the literature, and rank the publications in order of the relevance which the manually annotated scopes imply.

### C. *Curation Interface*

Both of these components will form a new curation interface to the HMMER web service. This will be available for use by Pfam's curators, and by users who have identified potentially new or improved Pfam families through their protein sequence similarity search query.

## Acknowledgement

We would like to thank the anonymous reviewers for their helpful comments.

## References

[1] Sonnhammer, E. L., Eddy, S. R., & Durbin, R. (1997). Pfam: a comprehensive database of protein domain families based on seed alignments. Proteins, 28(3), 405–20.

[2] Finn, R. D., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Mistry, J., Mitchell, A. L., Potter, S.C., Punta, M., Qureshi, M., Sangrador-Vegas, A., Salazar, G.A., Tate, J., Bateman, A. (2016). The Pfam protein families database: towards a more sustainable future. Nucleic Acids Research, 44(Database issue), D279–D279–85. doi:10.1093/nar/gkv1344

[3] Eddy, S. R. (2011). Accelerated Profile HMM Searches. PLoS Computational Biology, 7(10), e1002195. doi:10.1371/journal.pcbi.1002195

[4] Finn, R. D., Clements, J., & Eddy, S. R. (2011). HMMER web server: interactive sequence similarity searching. Nucleic Acids Research, 39(Web Server issue), W29–37. doi:10.1093/nar/gkr367

[5] Punta, M., Coggill, P. C., Eberhardt, R. Y., Mistry, J., Tate, J., Boursnell, C., Pang, N., Forslund, K., Ceric, G., Clements, J., Heger, A., Holm, L., Sonnhammer, E.L., Eddy, S.R., Bateman, A., Finn, R. D. (2012). The Pfam protein families database. Nucleic Acids Research, 40(Database issue), D290–301. doi:10.1093/nar/gkr1065

[6] Gardner, P. P., Daub, J., Tate, J., Moore, B. L., Osuch, I. H., Griffiths-Jones, S., Finn R.D., Nawrocki, E.P., Kolbe, D.L., Eddy, S.R., Bateman, A. (2011). Rfam: Wikipedia, clans and the “decimal” release. Nucleic Acids Research, 39(Database issue), D141–5. doi:10.1093/nar/gkq1129

[7] Krallinger, M., Valencia, A., & Hirschman, L. (2008). Linking genes to literature: text mining, information extraction, and retrieval applications for biology. Genome Biology, 9 Suppl 2(Suppl 2), S8. doi:10.1186/gb-2008-9-s2-s8

[8] Jaroszewski, L., Koska, L., Sedova, M., & Godzik, A. (2014). PubServer: literature searches by homology. *Nucleic Acids Research*, *42*(Web Server issue), W430-5. doi: 10.1093/nar/gku450

[9] The UniProt Consortium. (2014). UniProt: a hub for protein information. Nucleic Acids Research, 43(D1), D204–212. doi:10.1093/nar/gku989

[10] Leaman, R., & Gonzalez, G. (2008). BANNER: an executable survey of advances in biomedical named entity recognition. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 652–63. PMID:18229723

[11] Cucerzan, S. (2007). Large-Scale Named Entity Disambiguation Based on Wikipedia Data. In Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL) (pp. 708–716). Prague, Czech Republic: Association for Computational Linguistic