Crowdsourcing Protein Family Database Curation

Matt Jeffryes, Alex Bateman
European Bioinformatics Institute (EMBL-EBI), Cambridge, United Kingdom
{mjj, agb}@ebi.ac.uk

1. INTRODUCTION

Protein families are groupings of proteins with an evolutionary relationship. Pfam is a protein family database which has been maintained since 1996. Protein families are described by an alignment which is constructed by curation staff. Pfam has coverage of 47% of the residues in *pfamseq*, the protein sequence database on which it is based.

The alignments which describe protein families are transformed into hidden Markov models (HMMs) by the software package HMMER. These HMMs are used to query *pfamseq*. Significant matches for this query form the members of the family.

HMMER can also be used to discover families. It is capable of sensitive sequence similarity search, and can find homologous proteins which are distantly related to the query sequence. When performing searches, HMMER constructs a HMM from the query sequence, and uses this to query the target sequence database, similar to the method used to identify members of Pfam families. This makes HMMER a useful tool for identifying new families.

In addition to the HMMER software package, the HMMER web service is hosted at the European Bioinformatics Institute. This enables users to search against a number of sequence databases, without having to acquire their own copies of the databases.

When constructing a family, curators search the scientific literature for evidence of its existence, and the possible functions and structures of members of the family.

We aim to introduce crowdsourcing into the creation of Pfam families, and to facilitate easier annotation of families, through the use of text mining.

2. METHOD

HMMER can be used to identify new families. It is used regularly for this purpose by Pfam curators. It can also occur that a user performs a sequence similarity search which matches a grouping of proteins which is not described by an existing Pfam family.

Adding such a grouping to Pfam would improve the coverage of the database, but presently, this will only happen if the user performing the search takes the initiative to submit it to Pfam through a manual email process. We propose that users of the HMMER web service should be alerted to the situation, and given the opportunity to submit such searches to Pfam.

We have implemented a prototype version of this system. The prototype can perform searches using the HMMER web service API, and is able to compare the resulting grouping of proteins to the existing families in Pfam. If the grouping is novel, the user is able to submit it to Pfam using a form email.

3. FUTURE WORK

In addition to the grouping of proteins which forms a family, a useful Pfam entry will be annotated with a detailed description of the family, and a summary of research into its function and structure. We would like automatically identify literature relevant to a particular family through the fusion of sequence similarity search and literature search. We are developing a classifier to identify literature

which discusses the function and structure of proteins, as we believe such literature is of the most use to curators.

We intend to augment HMMER web service sequence similarity search results with relevant literature identified by this classifier. This will facilitate rapid annotation of newly identified families. We intend that this tool would be used for Pfam curators, and by users who identify possible new Pfam families.

The automatic identification of novel families, and the literature search would form an integrated tool for the crowdsourcing of new Pfam families, and their annotation.