Quality Control in the Pfam Database

Alex Bateman, Ewan Birney, Kevin Howe, Richard Durbin.

The Sanger Centre, Wellcome Trust Genome Campus, CB10 1SA. UK.

E-mail: agb@sanger.ac.uk

Pfam is a collection of protein domain family alignments, hidden Markov models and documentation. Pfam is composed of two components: Pfam-A and Pfam-B. Pfam-A is the curated part of Pfam that contains over 1400 families. Pfam-B is a collection of automatically generated domain families, derived from Prodom, that are not curated.

Keeping the large number of curated entries up to date and at a high quality presents a number of difficulties. Pfam attempts to give biologically useful annotation for each family. The annotation in Pfam has been added by a variety of authors. To ensure that the annotation is consistent we use a variety of quality control scripts to enforce the Pfam format.

Within the Interpro partnership we are attempting to merge our annotation with other domain databases. This has required us to alter some of the structure of our annotation to allow the functional comments to be parsed automatically.

Pfam-B is now a collaboration between Pfam and our Interpro partner Prodom. The latest releases of Pfam-B are derived from the Prodom database. This collaboration has given us a new quality control that shows where Prodom and Pfam-A disagree. This has shown in many cases where a Pfam-A family can be extended to include new members. The collagen family is used as an example.

Probably the single most important quality control is the overlap criterion. This says that no single residue can belong to more than one protein family. This allows us to spot false positives in families rapidly and spot where families can be merged. With the residue coverage of Pfam being over 40% this control is very strong. As the coverage in Pfam increases the overlap criterion becomes a stronger quality control.