5. Save a thousand words: Advanced visualisations for Pfam and iPfam

Benjamin Schuster-Böckler, Robert D. Finn, Alex Bateman
Sanger Institute, Hinxton, UK

The Pfam database of protein domain families has grown steadily since 1997. New features have gradually been added to it. We present a classification of related families called Pfam clans and a database of domain interactions called iPfam together with visualisations to improve the accessibility and usability of these resources.

The Pfam database encompasses a large set of semi-automatically created protein domain families. Since its beginning, it has grown steadily and by now covers more than 75% of all known protein sequences. Together with the growth in data, new features were added to Pfam. These include links to proteins with known three-dimensional structure, a classification of related families called Pfam clans and a database of domain-domain interactions seen in crystal structures called iPfam. To improve accessibility and usability, we seek to employ user friendly visualisations. We will illustrate these advances in presenting Pfam related data and how this improves understanding of protein function.

In Pfam, domain families are represented as profile hidden Markov models(HMMs). These can be visualised with an HMM Logo, a diagram combining relative aminoacid frequencies, residue information content and insert/delete probabilities into one comprehensive chart. We extended this concept to visualise pairwise comparisons of domain families. We use such pairwise HMM-Alignments, created with Martin Madera's profile comparer PRC, to analyse the similarity of Pfam families. We calculate scores for each Pfam family against all other families. Pfam clans are then built for significant hits to capture distant homologies. To depict this network of relationships within a clan, we provide a graph where nodes are domain families and edges represent the E-Values from the HMM alignments. Clicking on an edge will link to a pairwise HMM Logo for the respective alignment. Such a pairwise HMM Logo consists of two HMM Logos aligned horizontally, every matching state pair being highlighted and connected by a bar.

iPfam is a database derived by mapping Pfam families onto PDB crystal structures. Domains seen interacting in a crystal structure are collected. A multiple alignment with structural markups of interacting residues for each PDB file that contains the selected domain interactions can be obtained. A graph depicting the architecture of interacting domains for each PDB entry is displayed. It distinguishes between inter and intrachain interactions and highlights the selected domain interactions.
Overall, these advances allow users of Pfam to better understand the relationships of proteins and how they interact with each other.