Clustering of proteins from SWISS-PROT and TrEMBL.

Evguenia Kriventseva, Wolfgang Fleischmann, Alexander Kanapin, Rolf Apweiler.
Evguenia.Kriventseva@ebi.ac.uk, fleischmann@ebi.ac.uk .

EMBL Outstation, European Bioinformatics Institute,
Wellcome Trust Genome Campus, Hinxton, Hinxton Hall,
Cambridge, CB10 1SD, United Kingdom.

We are developing a database of SWISS-PROT and TrEMBL protein clusters which is created in a fully automatic way. The clusters will be used for the automatic annotation and removing redundancy in TrEMBL as well as for the search of new protein families and work with multi-domain proteins.

The approach is based on two steps. Firstly, a similarity matrix of "all against all" protein sequences is built. The similarity is computed using the Smith-Waterman algorithm with a Monte-Carlo simulation to estimate the significance, resulting in a z-score. Secondly, the clusters are built for different levels of protein similarity.

For an update run we keep all scores of unchanged sequences and only compute "new against new" and "new against unchanged", which avoids time-consuming calculations.