How UniProtKB/TrEMBL Tackles High Redundant Proteomes

Benoît Bely and UniProt Consortium

1EMBL-European Bioinformatics Institute, Cambridge, UK
2EBI Swiss Institute of Bioinformatics, Geneva, Switzerland
3Protein Information Resource, Georgetown University, Washington DC & University of Delaware, USA

Abstract

The UniProt Knowledgebase (UniProtKB) is a comprehensive resource for protein sequences and their annotation. As a result of the vast increase in genome sequencing projects in the last few years, UniProtKB has doubled in size in the last year to nearly 90 million entries with a high level of redundancy. This is especially true for bacterial species where different strains of the same bacterium have been sequenced and submitted (e.g. 4,080 proteomes for Staphylococcus aureus comprising 10.88 M entries and 1,692 strains of Mycobacterium tuberculosis making up 6.38 M entries). To deal with this redundancy, we developed a procedure to identify highly redundant proteomes within species groups. This procedure has been implemented for bacterial species and the sequences corresponding to redundant proteomes (approximately 50 million entries) were deprecated. All proteomes remain searchable through UniProt’s Proteomes interface (http://www.uniprot.org/proteomes/) and redundant proteome sets are now available for download from the UniProt Archive (UniParc).

The method

1. Group proteomes by species
   - Proteomes can only be redundant to other proteomes of the same species
   - Further: Proteomes can only be redundant to other proteomes within the same subspecies branch

2. Pairwise comparison of proteomes
   - CD-Hit-2D
   - Proteome similarity = (2 Length of matched sequence × alignment similarity) / (Length of matched sequence + Length of unmatched sequence)

3. Ranking
   - Indegree: How many proteomes are redundant to proteome A in a given network (the higher the better)
   - Outdegree: To how many proteomes proteome A is redundant (the lower the better)
   - Example: Proteome A (Indegree, Outdegree) 4/6; 1/4 is worse than 6/4; 3/6 is worse than 4/4

4. Elimination of redundant proteomes
   - Allochromas macleodi
   - 7 strains
   - Information about each removed node must be preserved in indegree neighbours.

5. Performance
   - Move redundant proteomes to UniParc
   - Repeat procedure only for new proteomes imported in each cycle
   - Recalculate redundancy every six months

Results

Release 2015_07 (24th June 2015)

- Total bacterial proteomes: 24963
  - Non-redundant: 8735
  - Redundant: 16228
    - Staphylococcus aureus: 4,080 proteomes ⇒ 21 non-redundant
    - Mycobacterium tuberculosis: 1,052 proteomes ⇒ 12 non-redundant

Funding

UniProt is funded by the European Molecular Biology Laboratory, National Institutes of Health, European Union, Swiss Federal Government, British Heart Foundation and National Science Foundation.

URL: www.uniprot.org