How UniProtKB/TrEMBL Tackles High Redundant Proteomes.

Benoît Bely  
UniProt release production project leader  
bbely@ebi.ac.uk
Introduction

Concept of proteome data growing very fast
Proteome portal: introduction

Welcome to the new UniProt website! We hope you enjoy the new design. If you're not quite ready yet, you can still go back to the old site.

The mission of UniProt is to provide the scientific community with a comprehensive, high-quality and freely accessible resource of protein sequence and functional information.

- UniProtKB
  - Swiss-Prot (546,790)
    - Manually annotated and reviewed.
  - TrEMBL (86,536,393)
    - Automatically annotated and not reviewed.

- UniRef
  - Sequence clusters

- UniParc
  - Sequence archive

- Proteomes

Supporting data

- Literature citations
- Taxonomy
- Subcellular locations
- Cross-ref. databases
- Diseases
- Keywords

News

K like Koagulation | Change of the cross-reference ArrayExpress to ExpressionAtlas
UniProt release 2014_10

Small is beautiful (and useful) | Evidences in the UniProtKB flat file format
UniProt release 2014_09

Ubiquitin caught at its own game | New

News archive
The Tsunami of data

2014_06 : 13 millions added
2014_07 : 11 millions added
Usual rate is 1.5-2 millions
=> Equivalent to 1 year of data
High number of genomes of same species

- *Staphylococcus aureus*, 3,599 proteomes -> 9.61 M entries
- *Mycobacterium tuberculosis*, 1,513 proteomes -> 5.97M entries
- *Acinetobacter baumannii*, 414 proteomes -> 1.61M entries
- *Salmonella enterica*, 184 proteomes -> 0.75M entries
Issues and consequences

• Redundancy of sequences in UniProtKB/TrEMBL
  • slowness in database searches
  • higher computational costs
  • increasing bias in statistical analyses
  • highly repetitive search results for over-represented sequence

• FTP repository
  • Difficult for user to use simple method to handle network disconnection and corrupted files, complete UniProtKB tarball
    • 2012_08 11Go
    • 2013_08 17Go
    • 2014_08 38Go
Solution
Proteomes redundancy removal

- Limit the number of alternate proteomes at the species level
- Identify redundant proteomes by performing pairwise comparison/alignments of sets of sequences for pairs of proteomes.
Method in 2 stages

• 1\textsuperscript{st} stage : Compare
  • Comparison module compares pairs of proteomes
  • Identify and weight proteomes similarity
• 2\textsuperscript{nd} stage : Analyse
  • Comparison data are used to build a graph
  • Reduction of the graph provide redundant and non-redundant proteomes
1st Comparison: Principle

- **CD-Hit-2D**
  - Compares two sets of sequences
  - *Sequence identity threshold 90%*
  - Proteome similarity threshold – 90%

\[
\text{Similarity} = \frac{\left(\sum \text{aligned sequence length} \times \text{alignment similarity}\right)}{\left(\sum \text{length of unmatched sequenced}\right)}
\]
1st Comparison : Heuristics

• Compare proteomes which are the leaves of a same taxonomy species branch.
  • *Proteomes can only be redundant to other proteomes of the same species.*

• Compare proteomes that contain a comparable number of proteins
1st Comparison: directionality

- A can be redundant to B
- Opposite does not need to be true
- Unidirectional
2nd Analysis module

- Redundancy is not transitive
- Data collected by comparison module can be presented in a form of a directed weighted graph.

Peptoclostridium difficile, 207 nodes
Dominating set for a graph $G = (V, E)$ is a subset $D$ of $V$ such that every vertex not in $D$ is adjacent to at least one member of $D$.

The domination number $\gamma(G)$ is the number of vertices in a smallest dominating set for $G$. 

(a) 

(b) 

(c)
2nd Analysis: Algorithm

- Iterative reduction
- Ranking
  - Indegree - how many proteomes are redundant to proteome A
  - Outdegree - to how many proteomes proteome A is redundant
  - Proteome importance score (Reference, published….)
  - Annotation level - number of Swiss-Prot Entries, annotation score
- Information weight — preserve information about removed nodes
- Mandatory (untouchable) proteomes — reintroduce at the end
1st result September 2014

- 49,588,619 Uniprot/TrEMBL entries in redundant proteomes

<table>
<thead>
<tr>
<th>Superregnum</th>
<th>Uniprot entries</th>
</tr>
</thead>
<tbody>
<tr>
<td>archaea</td>
<td>41,069</td>
</tr>
<tr>
<td>bacteria</td>
<td>49,222,850</td>
</tr>
<tr>
<td>eukaryota</td>
<td>319,602</td>
</tr>
<tr>
<td>viruses</td>
<td>5,098</td>
</tr>
</tbody>
</table>
Proteome redundancy removal

Applied for very 1st time for UniProtKB release 2015_04
Applied to Bacteria only

- **Staphylococcus aureus**: 4,080 proteomes → 21 non-redundant
- **Mycobacterium tuberculosis**: 1,692 proteomes → 12 non-redundant

~50 million UniProtKB/TrEMBL entries

(35.54%) non-redundant

(64.46%) redundant
Number of entries in UniProtKB/TrEMBL
UniProtKB taxonomy distribution

2015_03
- 14.68% Bacteria
- 2.40% Archaea
- 0.60% Eukaryota
- 0.99% Viruses
- 81.33% Other

2015_04
- 62.29% Bacteria
- 4.79% Archaea
- 1.19% Eukaryota
- 1.96% Viruses
- 29.78% Other
How deleted UniProtKB records look like
How to retrieve redundant proteomes

Alteromonas macleodii AltDE1

- This proteome is redundant to UP000014905.

Overview

- Proteins: 4,263 sequences in UniParc
- Proteome ID: UP0000000222
- Taxonomy: 1004786 - Alteromonas macleodii AltDE1
- Last modified: February 12, 2015
- Genome assembly: GCA_000310085.1

Components

- Plasmid pAMDE1: CP003918
- Chromosome: CP003917

UniParc sequence(s)

- 251
- 4012
Acknowledgment

EBI:
Borisas Bursteinas
Ramona Britto

Alan Wilter Sousa da Silva
Benoît Bely
Claire O’donovan
Hermann Zellner
Jie Luo
Maria Martin
Maurizio De Giorgi
Steven Rosanoff

SIB:
Andrea Auchincloss
Catherine Rivoire
Ivo Pedruzzi

Elisabeth Gasteiger
Emmanuel Boutet
Jerven Bolleman
Parit Bansal
Nicole Redaschi
Severine Duvaud
Sylvain Poux

PIR:
Chuming Chen
Hongzhan Huang
Peter McGarvey
1st Comparison: Similarity example (fake)

- A: Enterobacteria phage MS2 – 5 proteins
- B: Enterobacteria phage MS2 (isolate ST4) – 4 proteins
- Group 1: length 545, similarity 98.53%
- Group 2, length 393, similarity, 99.75%
- Group 3, length 130, similarity 100%
- Group 4, length 75, similarity 100%
- Group 5, length 45, similarity 0%

\[
\text{Similarity} = \frac{(545 \times 0.9853 + 393 \times 0.9975 + 130 \times 1.0 + 75 \times 1.0 + 45 \times 0)}{(545 + 393 + 130 + 75 + 45)}
\]

\[
= 95.45\%
\]
1st Comparison: Performance

- 1500-2000 comparisons per hour
- Runs daily
- Comparison stats
  - All proteome comparisons: 630,045
  - All qualified proteome comparisons: 297,142
2nd Analysis: Example
2nd Analysis: Example
2nd Analysis : Example

- 'Ionian Sea UM44b'
  1.0
  A

- 'Aegean Sea MED64'
  0.2
  AB

- 'Ionian Sea U4'
  2.2
  AB

- 'Ionian Sea U8'
  3.2
  A

- 'Ionian Sea U7'
  2.2
  AB
2nd Analysis: Example

'Ionian Sea U4'
2,2
AB

'Ionian Sea U7'
2,2
AB

'Ionian Sea U8'
2,2
AC

'Ionian Sea UM4b'
0,0
ABC
2nd Analysis : Example

'Ionian Sea UM4b'
0,0
A
B
C

'Ionian Sea U4'
2,2
A
B
D
2nd Analysis: Example