databases and protein inference

lennart martens

lennart.martens@vib-ugent.be
computational omics and systems biology group
VIB / Ghent University, Ghent, Belgium
Sequence Redundancy and Peptides

Enriching Sequence Databases

- DBToolkit
- Why Processing Matters

Time Lability of Sequence Databases

- The PICR Mapping Service

Protein Inference

- Algorithms for the Protein Inference Problem
- Protein Inference and Quantification
Sequence Redundancy and Peptides

Enriching Sequence Databases
- DBToolkit
- Why Processing Matters

Time Lability of Sequence Databases
- The PICR Mapping Service

Protein Inference
- Algorithms for the Protein Inference Problem
- Protein Inference and Quantification
Peptide-level sequence redundancy

>Protein 1
LENNARTMARTENS
>Protein 2
LENNARTMARTENT

non-redundant protein DB ≠ non-redundant peptide DB

Database content: all peptide sequences in the database
Database information: number of unique peptide sequences
Database information ratio: \[
\frac{\text{database information}}{\text{database content}}
\]
Tryptic cleavage, 1 allowed missed cleavage, Mass limits from 600 to 4000 Da.

Information ratios for common databases

See: Barsnes and Martens, Amino Acids, 2013
Sequence Redundancy and Peptides

**Enriching Sequence Databases**

*DBToolkit*

*Why Processing Matters*

**Time Lability of Sequence Databases**

*The PICR Mapping Service*

**Protein Inference**

*Algorithms for the Protein Inference Problem*

*Protein Inference and Quantification*
The influence of the sequence database

In vivo processing

Enzymatic digest and subsequent NH$_2$-terminal peptide isolation

Not in the sequence database!
Mitochondrial Isovaleryl-coA Dehydrogenase

**N-terminal transit peptide (1-29)**

MATATRLLGWRVASWRLRPPLLAGFVS

**Isovaleryl-CoA dehydrogenase (30 – 423)**

QRAHSLLPVDADAINGLSEEQQLRE...

...LDGIQCFGGGNGYINDFPMGFRFLRDA

KLYEIGAGTSEVRRLVIGRAFNAFDFH

An example
Extending the information content

AHSLLPVDDAINGLSEEQR → AHSLLPVDDAINGLSEEQR

HSLLPVDDAINGLSEEQR
SLLPVDDAINGLSEEQR
LLPVDDAINGLSEEQR
LPVDDAINGLSEEQR
PVDDAINGLSEEQR
VDDAINGLSEEQR

......
Another example: *in vivo* protein cleavage

Caspase cleavage of this protein (for 50%)

NH\textsubscript{2}-terminal peptide isolation

NOT IN DB!
Solving the issue: *in silico* bifunctional enzymes

The **bifunctional enzyme** generates the correct peptides!

**Title:** Arg-C  
**Cleavage:** R  
**Restrict:** P  
**Cterm**

**Arg-C definition**

---

**Title:** dualArgC_Cathep  
**Cleavage:** DXR  
**Restrict:** P  
**Cterm**

**Arg-C (N-term), Cathepsin (C-term) definition**
Sequence Redundancy and Peptides

Enriching Sequence Databases

**DBToolkit**

*Why Processing Matters*

Time Lability of Sequence Databases

*The PICR Mapping Service*

Protein Inference

*Algorithms for the Protein Inference Problem*

*Protein Inference and Quantification*
Working with databases: DBToolkit

http://genesis.UGent.be/dbtoolkit

See: Martens et al., Bioinformatics 2005
Summary of DBToolkit functionalities

a) Enzymatic digestion using regular or ‘dual’ enzymes
   → proteins to peptides

b) N-terminal or C-terminal ragging
   → enhancing the information content of the database

c) Non-lossy redundancy clearing
   → raising database information ratio

d) Create shuffled and reversed databases
   → false-positives testing

e) Extract sequence-based subsets
   → a priori prediction of potential success rate

f) Map peptides back to proteins (maximal annotation approach)
   → find all matching proteins, and select primaries

etc ...

See: Martens et al., Bioinformatics 2005
Sequence Redundancy and Peptides

Enriching Sequence Databases

DBToolkit

Why Processing Matters

Time Lability of Sequence Databases

The PICR Mapping Service

Protein Inference

Algorithms for the Protein Inference Problem

Protein Inference and Quantification
Serum degradation over time

From: Yi et al., Journal of Proteome Research 2007
Plasma degradation over time

From: Yi et al., Journal of Proteome Research 2007
Aberrant processing as biomarker

(A) Plasma kallikrein

Heavy chain: procoagulant activity
Light chain: serine protease

Normal status
H1 H2 H3 H4 + Light chain

Lung cancer patients
H1 H2 H3 + Light chain
H4

(B) 17 kDa KLKB1 fragments

Normal Lung adenocarcinoma

Sequence Redundancy and Peptides

Enriching Sequence Databases

DBToolkit
Why Processing Matters

Time Lability of Sequence Databases

The PICR Mapping Service

Protein Inference

Algorithms for the Protein Inference Problem
Protein Inference and Quantification
## Example 1: HUPO PPP actualisation

**Bringing the PPP from IPI 2.21 to IPI 3.13**

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1555</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Changed</td>
<td></td>
<td></td>
</tr>
<tr>
<td>507</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unchanged</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1048</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Of which:**

- 338 Propagated 22%  67% (of ‘Changed’)
- 169 Defunct 11% 33% (of ‘Changed’)

**Of which**

- 95 Defunct (RFSQ_XP)  6%  56% (of ‘Defunct’)
- 72 Defunct (Ensembl)  5%  43% (of ‘Defunct’)
- 2 UniProt 0% 1% (of ‘Defunct’)

Both exist, 1 taxonomy now: **RAT**

1 immunoglobin

1048 + 345 = **1386** recoverable (89.1%)

*See: Martens and Mueller et al., Proteomics 2006*
Example 2: human blood platelets

**Bringing the Platelets from IPI 2.31 to IPI 3.13**

- Total: 673
- Unchanged: 578 (86%)
- Changed: 95 (14%)

**Of which:**

- Propagated: 78 (12%) 82% (of ‘Changed’)
- Defunct: 17 (3%) 18% (of ‘Changed’)

**Of which**

- Defunct (RFSQ_XP): 5 (1%) 29% (of ‘Defunct’)
- Defunct (Ensembl): 12 (2%) 71% (of ‘Defunct’)

578 + 78 = 656 recoverable (97%)

*See: Martens and Mueller et al., Proteomics 2006*
Proteins sometimes age badly
Sequence Redundancy and Peptides

Enriching Sequence Databases

*DBToolkit*

*Why Processing Matters*

Time Lability of Sequence Databases

*The PICR Mapping Service*

Protein Inference

*Algorithms for the Protein Inference Problem*

*Protein Inference and Quantification*
Submit accessions OR sequences (FASTA) with 500 entry interactive limit (no batch limit)

Select output format

Limit search by taxonomy (pessimistic)

Choose to return all mappings or only active ones

Select one or many databases to map to in one request

Run search

See: Côté et al., BMC Bioinformatics 2007

http://www.ebi.ac.uk/tools/picr
## Mapping results

Your search is in progress:

This page will be updated automatically when your search is done. If you want to refresh it now, click here.

### PICR - Protein Identifier Cross-Reference Service [BETA]

<table>
<thead>
<tr>
<th>Input Accession</th>
<th>SWISSPROT</th>
<th>TREMBL</th>
</tr>
</thead>
<tbody>
<tr>
<td>15079369</td>
<td>015526</td>
<td>O00273.1</td>
</tr>
<tr>
<td>015526</td>
<td>015526.1</td>
<td>O00273.1</td>
</tr>
<tr>
<td>RYBA2</td>
<td>O2PA4A 1</td>
<td>O00273.1</td>
</tr>
<tr>
<td>21399361</td>
<td>O06B45.1</td>
<td>O00273.1</td>
</tr>
<tr>
<td>17149328</td>
<td>O00273.1</td>
<td>O00273.1</td>
</tr>
<tr>
<td>O00273</td>
<td>O00273.1</td>
<td>O06B45.1</td>
</tr>
<tr>
<td>O00273</td>
<td>O00273.1</td>
<td>O06B45.1</td>
</tr>
<tr>
<td>21399362</td>
<td>O06B45.1</td>
<td>O06B45.1</td>
</tr>
<tr>
<td>P04901</td>
<td>P04901.1</td>
<td>O05058.1</td>
</tr>
<tr>
<td>P43251</td>
<td>P04901.1</td>
<td>O05058.1</td>
</tr>
<tr>
<td>Q9H5N1</td>
<td>Q9H5N1.1</td>
<td>O9H5N1.2</td>
</tr>
<tr>
<td></td>
<td>Q9H5N1.2</td>
<td>O9H5N1.2</td>
</tr>
<tr>
<td></td>
<td>Q9H5N1.2</td>
<td>O9H5N1.2</td>
</tr>
</tbody>
</table>

You can download your results as a [CSV file](#) or as a [XLS file](#) or click here to start another search.
Sequence Redundancy and Peptides

Enriching Sequence Databases
  DBToolkit
  Why Processing Matters

Time Lability of Sequence Databases
  The PICR Mapping Service

Protein Inference
  Algorithms for the Protein Inference Problem
  Protein Inference and Quantification
Not all peptides are created equal

Gene

Transcripts

Translations

Peptides

matching all transcripts
matching a transcript subset
matching exactly 1 translation
Sample preparation consequences

See: Nesvizhskii Al et al, Molecular and Cellular Proteomics, 2005
Sample preparation consequences

See: Nesvizhskii AI et al, Molecular and Cellular Proteomics, 2005
Protein inference: a question of conviction

<table>
<thead>
<tr>
<th>peptides</th>
<th>a</th>
<th>b</th>
<th>c</th>
<th>d</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>proteins</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>prot X</td>
<td>x</td>
<td></td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>prot Y</td>
<td></td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>prot Z</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
</tbody>
</table>

**Minimal set**

**Occam**

**Maximal set**

**anti-Occam**

**Minimal set with maximal annotation**

**true Occam?**

See: Martens and Hermjakob, Molecular BioSystems, 2007
Sequence Redundancy and Peptides

Enriching Sequence Databases

- DBToolkit
- Why Processing Matters

Time Lability of Sequence Databases

- The PICR Mapping Service

Protein Inference

- Algorithms for the Protein Inference Problem
A few algorithms for protein inference

- **IDPicker**
  
  *Zhang et al, Journal of Proteome Research, 2007*

- **ProteinProphet**
  
  *Nesvizhskii Al et al, Analytical Chemistry, 2003*

- **DBToolkit**
  
  *Martens et al, Bioinformatics, 2005*
IDPicker parsimonious protein assembly

(1) Initialize

IDPicker parsimonious protein assembly

(11) Collapse

(III) Separate

IDPicker parsimonious protein assembly

(IV) Reduce

In iteration 1, all weights $w$ start off as $1/n$, with $n$ the degeneracy count for the peptide.

See: Nesvizhskii AI et al., Analytical Chemistry, 2003
DBToolkit protein inference

Minimal set with maximal annotation

peptides

proteins

prot X (-)  x
prot Y (+)  x
prot Z (0)  x x x x

Accession Start  Stop  Previous Sequence Description/Sources

105 117  K  DTDDVFIM SWASS-FI  P0041255.1 (105-117)*API00796345.1 (105-117)*API00815192.1 (105-117)
148 167  R  DETMIVY SWASS-FI  P00000477.3 (129-141)
756 766  K  YEMFAQT SWASS-FI  P0081463.1 (756-766)*API00817771.1 (424-435)
428 441  R  SEQPTDS SWASS-FI  P00479106.1 (429-441)
1304 1613  R  TEMEDLM SWASS-FI  P00395722.4 (1306-1350)*API00478230.1 (107-116)
83 101  K  DOOEAAL SWASS-FI  P00793131.1 (83-101)
125 126  R  ELAILLG SWASS PROT P30361 [TREMBL: AUC16;8772] [ENSEMBL: ENSP0000342028] [REFSEQ: NP_04389] [H-INV: H0100023339]
261 301  R  VLAVNOES SWASS-FI  P00759776.1 (265-291)
21 30  K  AGFADGCSWASS-FI  P00014282.1 (21-30)*API00034011.1 (21-30)*API00032006.1 (21-30)*API00025416.1 (21-30)
159 189  R  DM13GEVE SWASS-FI  P10485[TREMBL: A38177;7856][ENSEMBL: ENSP0000248996][REFSEQ: NP_002364][H-INV: H0100023339]
105 106  R  OMESELE SWASS-FI  P00793131.1 (105-106)*API00793131.1 (105-106)*API00472341.1 (470-490)
490 491  K  DGVEMISWASS-FI  P00816336.1 (487-490)
2622 2631  R  DKGYTL SWASS-FI  P00302582.2 (2614-2625)*API00646467.1 (2552-2561)
119 121  K  THEADICSWASS-FI  P00816336.1 (119-121)*API00472341.1 (470-490)
198 198  K  MSLFYAE SWASS-FI  P00308384.2 (187-199)*API00472341.1 (107-113)*API00793131.1 (105-106)
197 200  K  ODLNMA SWASS-FI  P00793141.1 (173-196)
195 196  K  NAINVEQ SWASS-FI  P00374519.3 (154-163)*API00871386.1 (153-162)*API00895425.1 (154-163)
264 264  K  QSETENG SWASS-FI  P00478743.5 (263-264)*API00308384.2 (263-264)*API00374519.3 (263-264)*API00793131.1 (105-106)
593 593  K  SKRNIP SWASS-FI  P00495367.1 (581-593)*API00445274.4 (580-592)*API00472341.1 (470-490)
365 365  R  NAMGSLASWASS-FI  P00221304.2 (357-365)*API00473085.3 (357-365)*API00445274.4 (580-592)
130 130  K  YOEHILF SWASS-FI  P00478743.5 (130-130)*API00308384.2 (130-130)*API00374519.3 (130-130)*API00793131.1 (105-106)
210 210  K  ELLPLYSWASS-FI  P00478743.5 (210-210)*API00793131.1 (105-106)*API00472341.1 (470-490)
222 222  K  LDYDEGASWASS-FI  P00478743.5 (221-222)*API00472341.1 (470-490)
142 142  R  HMLASWASS-FI  P00374519.3 (142-142)*API00472341.1 (470-490)
187 187  K  QTLUNIQLEWASS-FI  P00374519.3 (187-187)*API00472341.1 (470-490)
266 266  K  XLMFCM1 SWASS-FI  P00478743.5 (266-266)*API00793131.1 (105-106)*API00472341.1 (470-490)
Some indications from the HUPO BPP

peptides

***proteins***

prot X (−)  x  x

prot Y (+)  x

prot Z (0)  x  x  x  x
Sequence Redundancy and Peptides

Enriching Sequence Databases
   *DBToolkit*
   *Why Processing Matters*

Time Lability of Sequence Databases
   *The PICR Mapping Service*

Protein Inference
   *Algorithms for the Protein Inference Problem*
   *Protein Inference and Quantification*
Some inference examples (i)

http://genesis.ugent.be/rover

Nice and easy, 1/1, only unique peptides (blue) and a narrow distribution

See: Colaert et al, Proteomics, 2010
Some inference examples (ii)

Nice and easy, down-regulated

See: Colaert et al, Proteomics, 2010
Some inference examples (iii)

A little less easy, up-regulated

See: Colaert et al, Proteomics, 2010
Some inference examples (iv)

A nice example of the mess of degenerate peptides

See: Colaert et al, Proteomics, 2010
Some inference examples (v)

A bit of chaos, but a defined core distribution

Thank you!

Questions?