Pathway Analysis
In Expression Proteomics

Roman Zubarev
Roman.Zubarev@ki.se

Molecular Biometry,
Department for Medical Biochemistry & Biophysics,
Karolinska Institutet, Stockholm
Proteomics vs Transcriptomics and Metabolomics

- **Genomics** - what the cell *may* do
- **Transcriptomics** - *wants* to do
- **Proteomics** - *does*
- **Metabolomics** - *has done*
Differences between transcriptomics and proteomics

- The dynamic range – $10^3$-$10^4$ vs $10^7$.

Since the dynamic range of instrumentation is $-10^3$-$10^4$, transcriptomics easily covers all 10,000 expressed genes, while proteomics – ca. 5,000 proteins. But false discovery rate for mRNA 5%, for proteins – 1%
Differences between transcriptomics and proteomics

- The cellular half-life:
  - mRNA - 9h
  - proteins - 46 h
Differences between transcriptomics and proteomics

- The number of protein molecules per mRNA: 1:1 to 1000:1
Combined Predictions - Length and AA Score

Log(Protein Translation Rate)

Predicted from Length and AA composition

Other factors contribute to translation rate!
mRNA abundances predict ca. 40% of the protein abundance, but log(Ratio) for mRNA predict >60% of log(Ratio) for proteins.

mRNA data need to be complemented by Proteomics data.
In three different cell lines, practically all expressed genes (and proteins) are shared.

Same proteins are expressed in every cell type, but with different abundances.
How does protein regulation depend upon protein abundance?

Protein regulation, arbitrary units

Protein abundance, log scale

No Difference

Experimental error

Full proteome
How does protein regulation depend upon protein abundance?

Small Difference

Protein abundance, log scale

Protein regulation, arbitrary units

Low abundant proteins - Relatively large fold-change

Highly abundant proteins - Relatively small fold-change
How does protein regulation depend upon protein abundance?

Large Difference

Relatively “flat” fold-change for the whole proteome

Protein regulation, arbitrary units

Protein abundance, log scale

Full proteome
SUMMARY

• Transcriptomics provides large (95%) coverage of expressed genes, but it explains, at best, only 40% of the log(Ratio) of protein abundances.

• Proteomics gives lower coverage (50% or less) by expressed proteins, but false discovery rate is only 1%

• For small changes in the proteome (e.g. early stages in time course), deep proteomics is advantageous, as proteins with significant fold-change are those of low-abundance

• For large changes in the proteome (e.g. cell type differentiation), even limited depth proteomics can provide specific fingerprint of cellular state, as protein regulation is largely independent upon abundance
Data Processing in Proteomics

Reductionist Molecular Biology:
- "golden bullet"
  - detailed interactions, modifications, mechanisms
  - lack of total picture

Statistical Approach:
- Ad hoc, empirical model
  - You get what you see
  - Prediction, accuracy
  - No explanation

Pathway Biology:
- Global model
  - prediction based on known pathways
  - unknown accuracy
  - do pathways exist?...
Protein Identification by Tandem Mass Spectrometry

Protein sequence
ILNKPEDETHLEAQPTDASAQFIRNLQISNELSKEPSISREDLISKEQIVIRSSRQPQSQNPKLPLSILKEKHLRNATLGSEETTEHTPSDASTTEGKLMELGHKIMRNLVENTKETIKYLKSLFShafevVkt

Enzymatic digest

Tryptic peptides
EDLISK
EQIVIR
LPLSILK
NLENTVK
LMELGHK
QPQSQNPKLNLQISNEDLSKNATLGSEETTEHTPSDASTTEGKSLFSHAFEVVK

Tandem Mass Spectrometry (MS/MS)

Tryptic peptide
NLENTVK

Fragmentation
MS/MS

Fragment masses
232.17
346.22
388.20
444.28
484.33
511.37
555.40
623.45
666.44
712.52

Molecular mass: 817.44

Your Peptide/protein is this:
Score = 77
“Deep” vs “Top” Proteomics

% of proteome coverage

2003 2004 2005 2006 2007 2008 2009 2010 2011
MS-based quantitative discovery approaches

**SILAC**  
- **Cells**: Heavy → Light  
- **Proteins**: H/L  
- **Peptides**: H/L

**TMT, iTRAQ**  
- **Cells**: Heavy  
- **Proteins**: H/L  
- **Peptides**: H/L

**Label-free**  
- **Cells**: N/A  
- **Proteins**: N≤8  
- **Peptides**: N-any

**Biology**  
- ?  
- 0  
- 0

**Extraction**  
- 0  
- x

**Digestion**  
- 0  
- x  
- x

**Tissues, bodily fluids**  
- N/A  
- N≤8  
- N-any

**Deep proteomics**  
-  
-  
- 

**Top proteomics**  
-  
-  
-  
Top Proteomics

- 'Top proteome': 1500-3000 proteins, 5000-9000 peptides
- No protein separation
- No peptide separation (on-line reverse-phase LC only)
- Single LC/MS experiment, 0.5-2.0 h long

≈4000 peptide identified, ≈1000 proteins quantified
Pathway Analysis

Disease Modeling

Patient Stratification

Drug Target Discovery

Establishing Drug Mechanism
Pathway Analysis & Proteomics
Analytical Pathway Biology

Receptor → Signal molecule → Adapters → Kinases → Transcription factors → mRNA → Proteome

BioBase - TRANSPATH database
GeneXplain - Analysis Tools
Pathway Analysis Workflow

1. Collect Sample Cells
2. Collect Control Cells
3. Perform Pathway Analysis
4. Analyze Pathways
5. Interpret Results
KeyNode-Mediated Analysis: Upstream

Pathway score: \( \sum (\text{keynode score}) \)

Score

KeyNode_1 3050
KeyNode_2 2987
KeyNode_3 2073
... KeyNode_N 25
DYNAMIC PROTEOMICS APPROACH
for drug target identification:
• by the speed of change (1 h), 10% selection
• by the total change in 48 h, 10% selection

Overall: top 3% (35 proteins)
Pathway Analysis of Dynamic Proteomics Data

I) Protein mapping on Pathways

Proteins from input list
Pathway Analysis of Dynamic Proteomics Data

Upstream Search:
• for Speed, 0-60 min
• for Magnitude, 0-2800 min

Key Nodes

KN Scoring: $\Delta S = (S_A - S_B) \times \log_2(S_A / S_B)$

Top KN is selected: one for Speed, one for Magnitude
The threshold problem in proteomics

\[ G = \text{Abs}(A_1 - A_2) \times \log_2(A_1 / A_2) \text{ [ppm]} \]

If statistical fluctuations of protein abundances follow Poisson distribution, \( G \)-threshold is constant

Pathway Analysis of Dynamic Proteomics Data

Downstream KN search

III) Two top KNs

Overlapping Molecules = Drug Target Candidates
Identification of TOPI as the drug target from 812 proteins in the input list

Overlap of downstream lists from $F_{\gamma}$, c-FLIP(h):
9 proteins, of which 2 from input list (known dynamics):

- TOPI, (speed + magnitude)-rank 228
- 26S proteasome, (speed+ magnitude)-rank 787
Overlap of downstream lists from $F_{\gamma}$, c-FLIP(h):
4 proteins, none from the input list:

- TOPI
- CKII
- Two NR-related proteins

What if TOPI is removed from Input list?..
Take-home messages:

- Transcriptomics and proteomics overlap, but proteomics is “closer to action”, and thus produces more relevant data.

- Proteomics is currently limited in “depth” due to the large dynamic range of protein abundances, but technology moves forward fast, and the proteomics depth is increasing.

- Correlation analysis provides first insight into the biological process, but pathway analysis is necessary to put the results in biological context.

- Simple mapping of regulated proteins onto pathways (“direct mapping”) often is insufficient.

- Upstream keynode analysis is superior over direct mapping.

- Combining transcriptomics, proteomics and metabolomics data is the future goal of pathway analysis.