Proteomics data analysis seminar

Quantitative proteomics and transcriptomics of anaerobic and aerobic yeast cultures reveals post-transcriptional regulation of key cellular processes

de Groot, M., Daran-Lapujade, P., van Breukelen, B., Knijnenburg, T., de Hulster, E., Reinders, M., Pronk, J., Heck, A. and Slijper, M.
Microbiology **153** (2007) 3864-3878

27.3. Paula Jouhten

Contents

- Quantitative proteomics background
- Experimental set up in the yeast case
- Generation of a robust data set
- Interpretation of the data
- Comparison to transcriptome data for revealing the regulatory level

Quantitative proteomics background

- Shortcomings in 2D gel based methods:

  poor reproducibility, biased for the most abundant proteins,...

- Mass spectrometry (MS) based quantitative proteomics

  MS is inherently not quantitative!

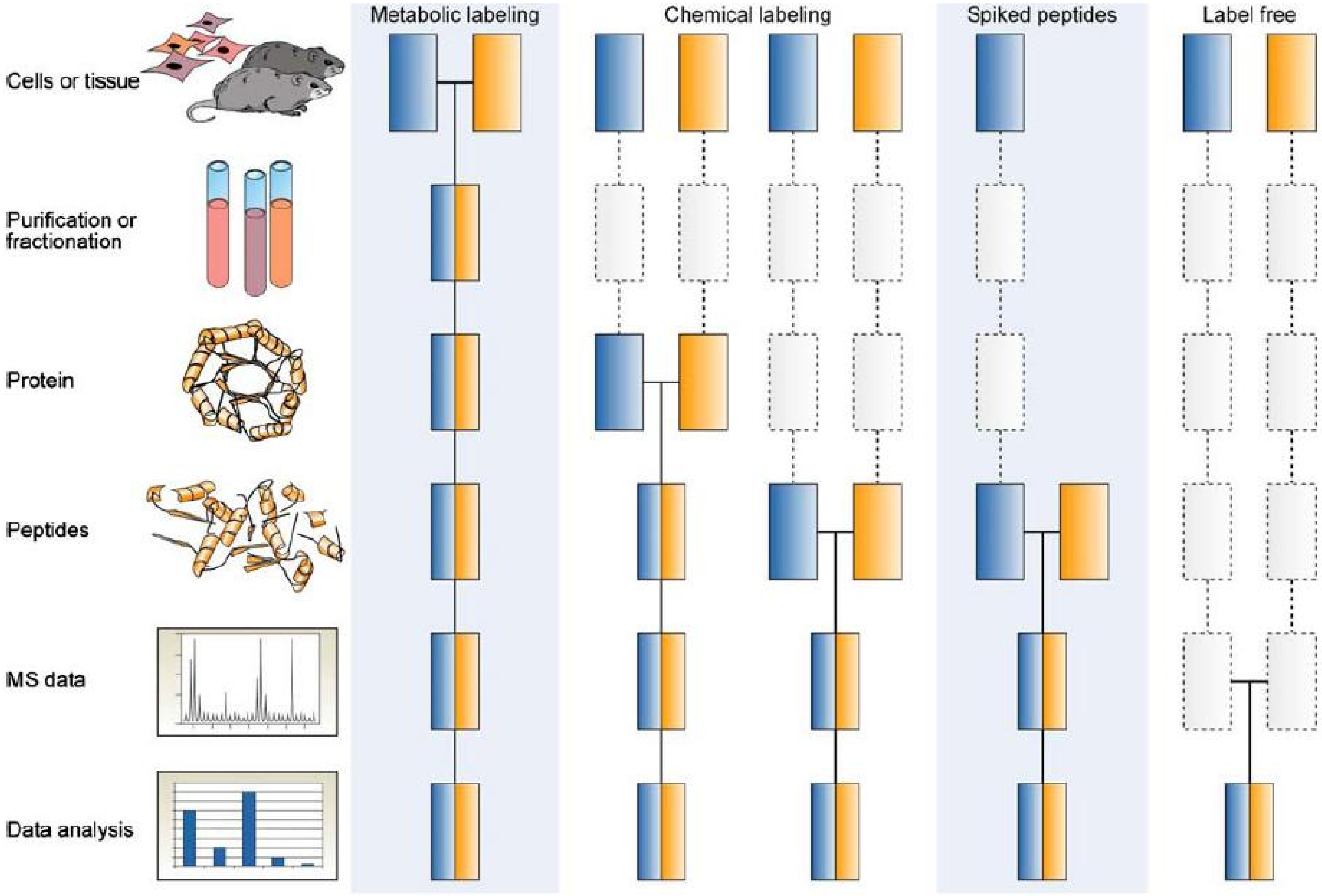
  physico-chemical properties affect the response

- Absolute vs relative quantification
- Mass tagging

  metabolic labelling
  isotope tagging
  enzymatic labelling
  labelled peptide standards

- Label-free quantification approaches

MS based quantification



Bantscheff *et al.*, *Anal Bioanal Chem* **389** (2007) 1017-1031.

Label-free approaches

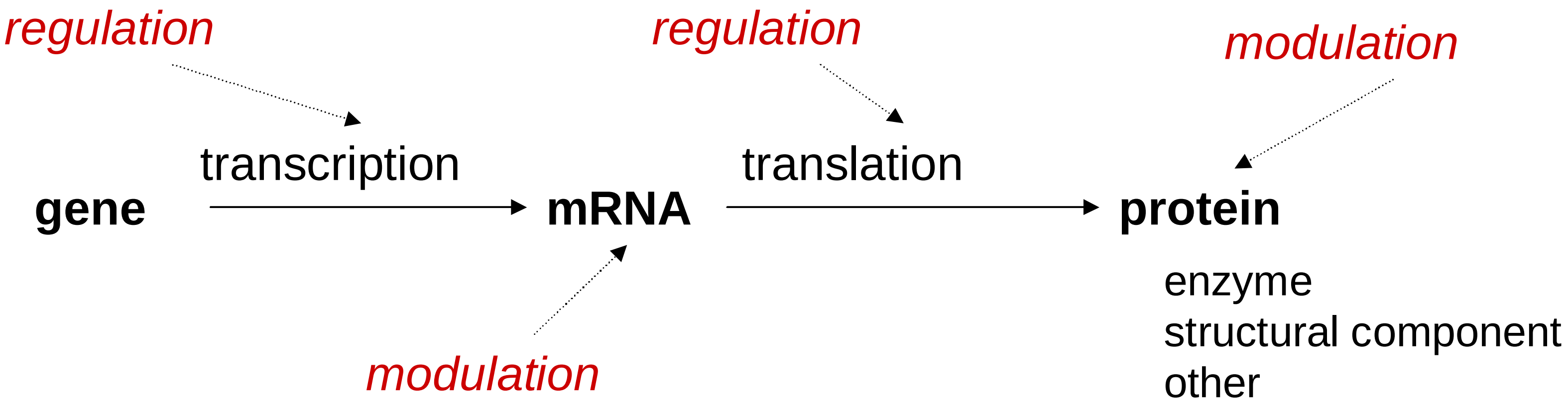
- Comparison of two or more experiments:

  1) comparison of direct MS signal intensity of any given peptide

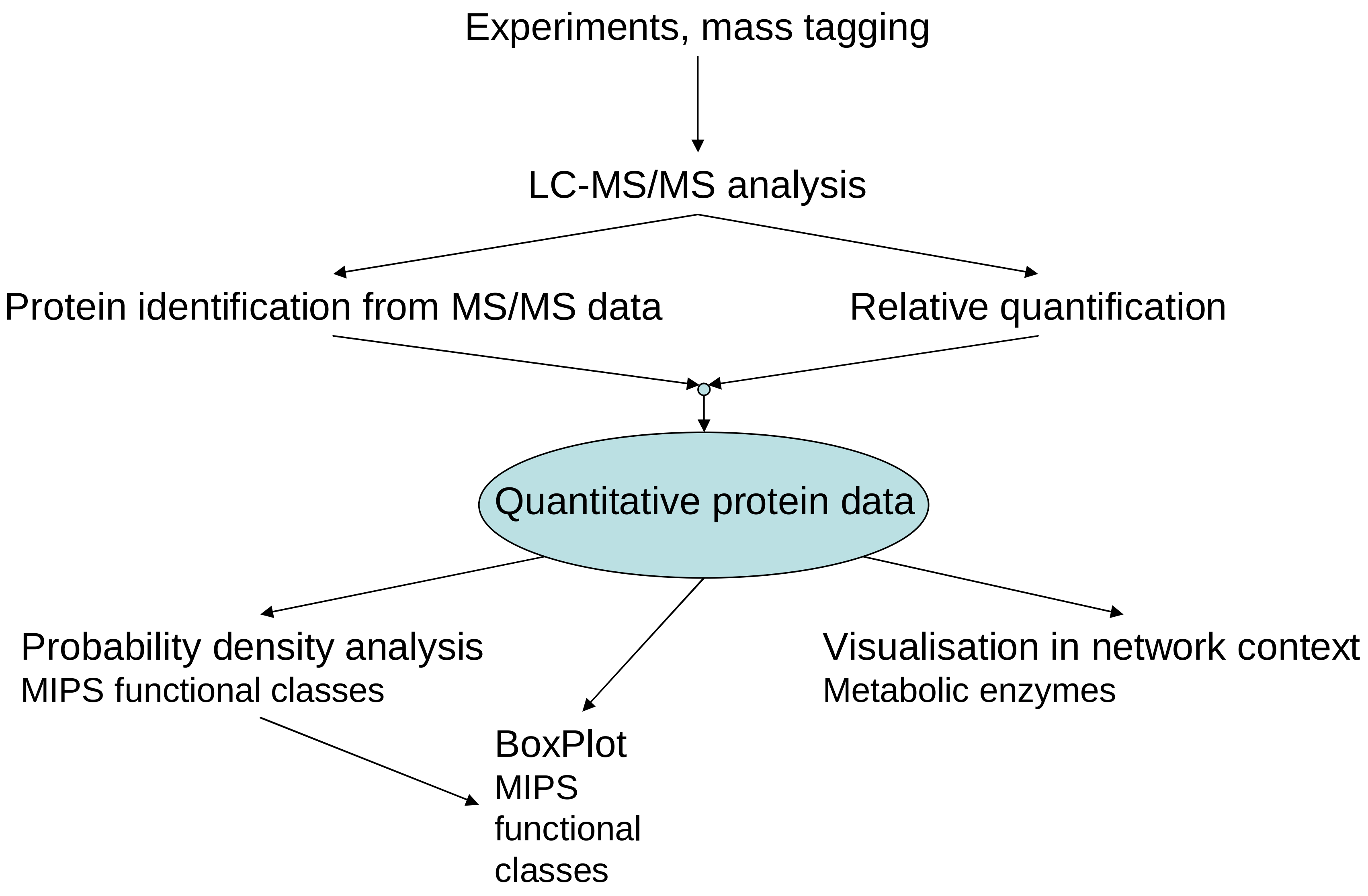
  2) comparison of a number of acquired fragment spectra matching to a peptide/protein = spectral counting

Quantitative proteomics and transcriptomics of anaerobic and aerobic yeast cultures reveals post-transcriptional regulation of key cellular processes

de Groot, M., Daran-Lapujade, P., van Breukelen, B., Knijnenburg, T.,
de Hulster, E., Reinders, M., Pronk, J., Heck, A. and Slijper, M.
Microbiology **153** (2007) 3864-3878



Workflow



Experiments, mass tagging

LC-MS/MS analysis

Protein identification from MS/MS data

Relative quantification

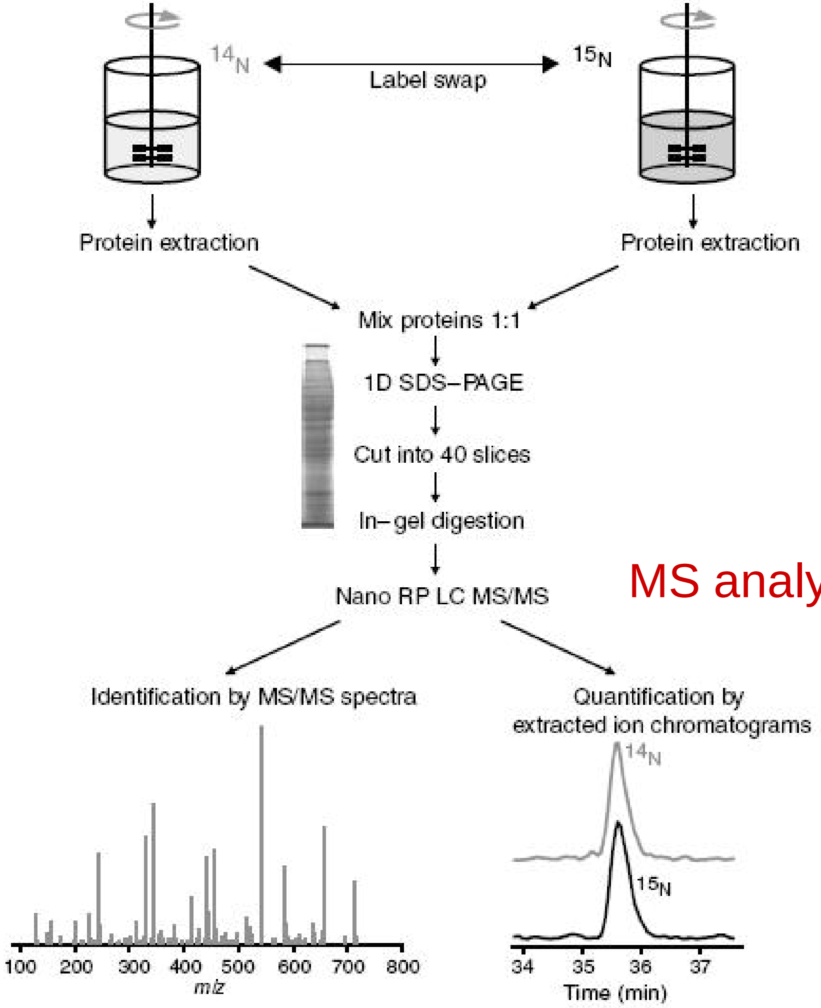
Quantitative protein data

Probability density analysis
MIPS functional classes

BoxPlot
MIPS
functional
classes

Visualisation in network context
Metabolic enzymes

Experimental set up



one biological replicate

MS analysis run in triplicate

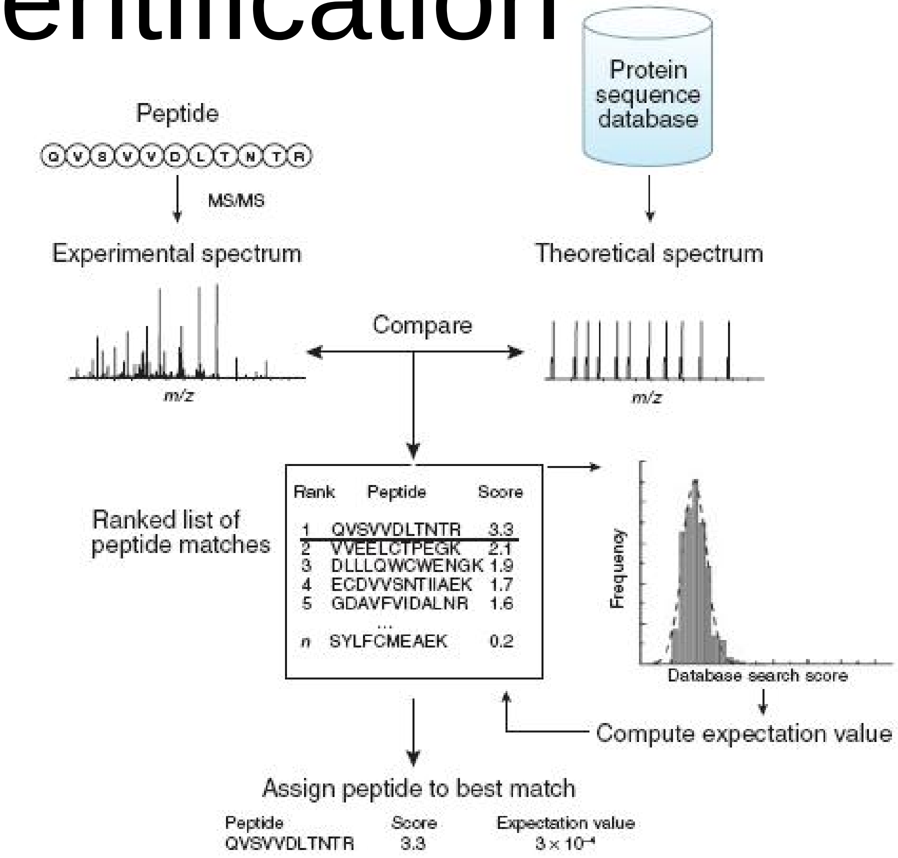
peptide data

Peptide identification

- SEQUEST for interpretation of MS/MS spectra against EBI proteome database, two runs for each data file (^{14}N, ^{15}N peptides)

  *each MS/MS spectrum individually
  *defines a set of candidate peptides with a matching mass from a DB
  *compares the experimental spectrum to the theoretical spectra

- Target-decoy searching for FDR


Protein sequence database
Peptide
QVSVVDLTNTR
MS/MS
Experimental spectrum
Theoretical spectrum
Compare
m/z
m/z
Ranked list of peptide matches
Frequency
Database search score
Compute expectation value
Assign peptide to best match

Rank	Peptide	Score
1	QVSVVDLTNTR	3.3
2	VVEELCTPEGK	2.1
3	DLLLQWCWENGK	1.9
4	ECDVVSNTIIAEK	1.7
5	GDAVFVIDALNR	1.6
	...	
n	SYLFCMEAEK	0.2

Peptide	Score	Expectation value
QVSVVDLTNTR	3.3	3×10^{-4}

Nesvizhskii *et al.*, *Nature Methods* **4** (2007) 787-797

Target-decoy search

- Search against the target DB
- Search against reversed (or randomised) DB
- Assumes the same distribution for matches to the decoy sequences and false matches to the original DB
- Other choice: empirical Bayes approaches

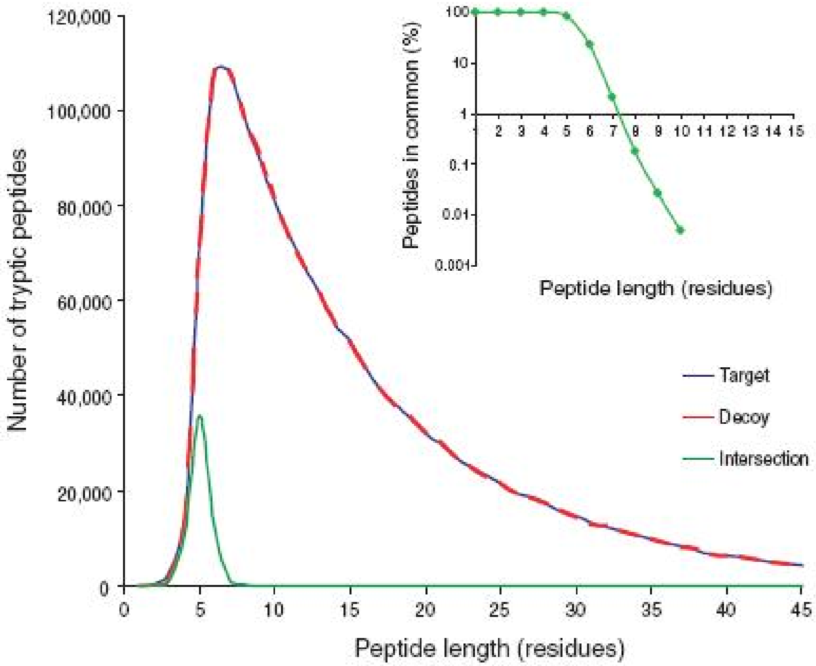


Figure 1 | Overlap between target (forward) and decoy (reversed) sequences is negligible. Human protein sequences within the minimally redundant International Protein Index sequence database[21] were digested *in silico* with trypsin (maximum two missed cleavage sites, maximum peptide length = 45; target). Tryptic peptides were similarly generated from the reversed protein sequences from this database (decoy). After converting isoleucines to leucines, the number of peptide sequences in common between the two databases was determined (intersection). Practically no peptides greater than 8 amino acids in length were found in both forward and reversed databases. Inset, percentage of peptides in common between target and decoy sequences decreases with increasing peptide length.

Elias and Gygi, *Nature Methods* **4** (2007) 207-214

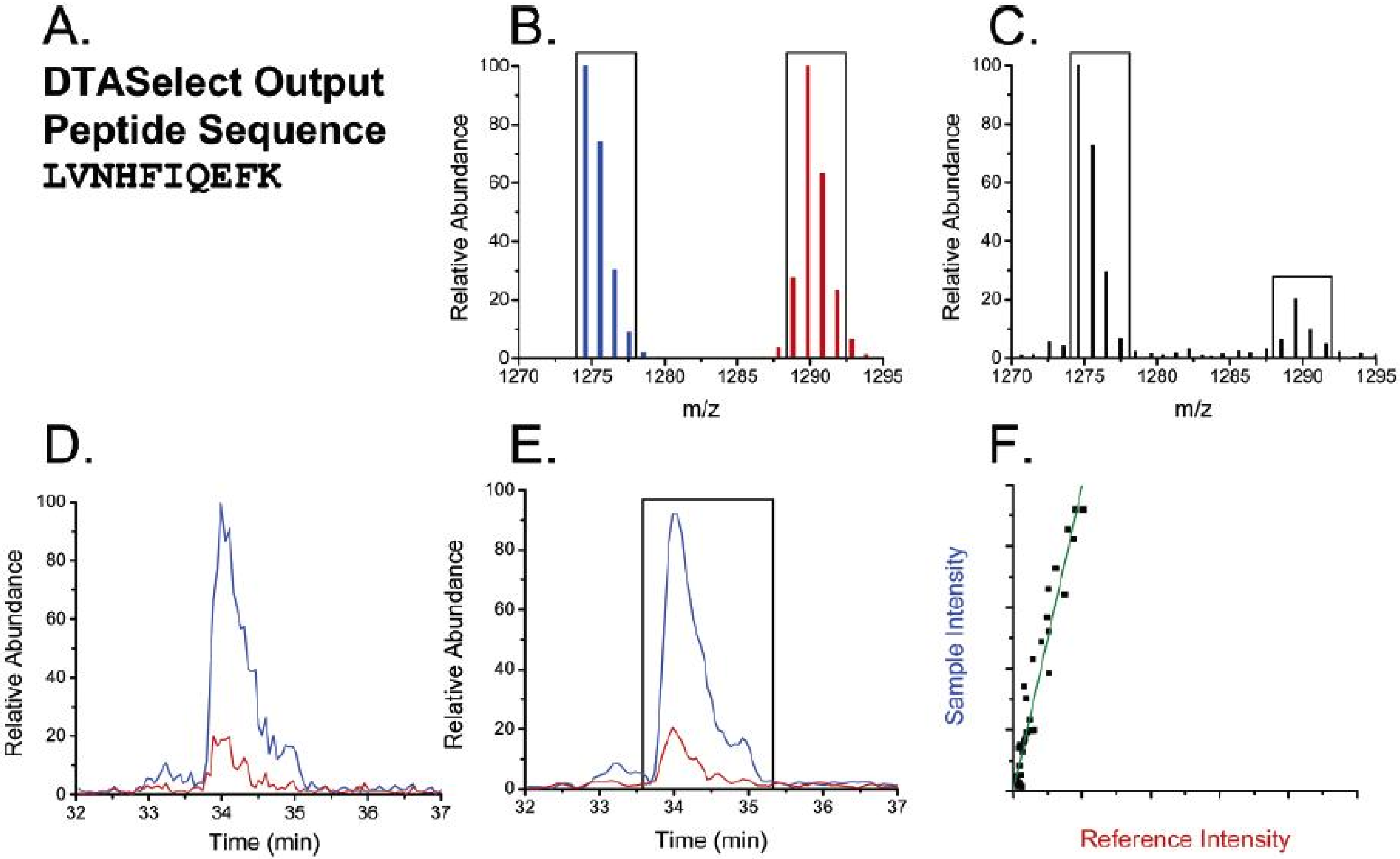
Protein identifications

- DTASelect for assembling the identified peptides into proteins
- SEQUEST output as input
- Sorts peptides by locus
- Full protein sequences again from the same DB
- User-defined criteria for selection of identifications
- ... FDR for protein identifications..?
- In the previous publication: minimum Xcorr 1.9, 2.2 and 3.75 for 1+, 2+, and 3+ peptides, respectively, and minimum deltaCNs 0.1 for each peptide

Relative quantification

- RelEx for calculation of peptide ion current ratios
  * Extraction of ion chromatograms
  * Smoothing
  * Peak detection
  * The peak nearest to the MS/MS spectrum is chosen for the calculation of the isotoper ratio
  * Linear least squares correlation
  * Sorting peptide ratios by protein locus
  * Omitting outliers (Dixon's Q-test)
  * Protein mean and std (t-test)

RelEx

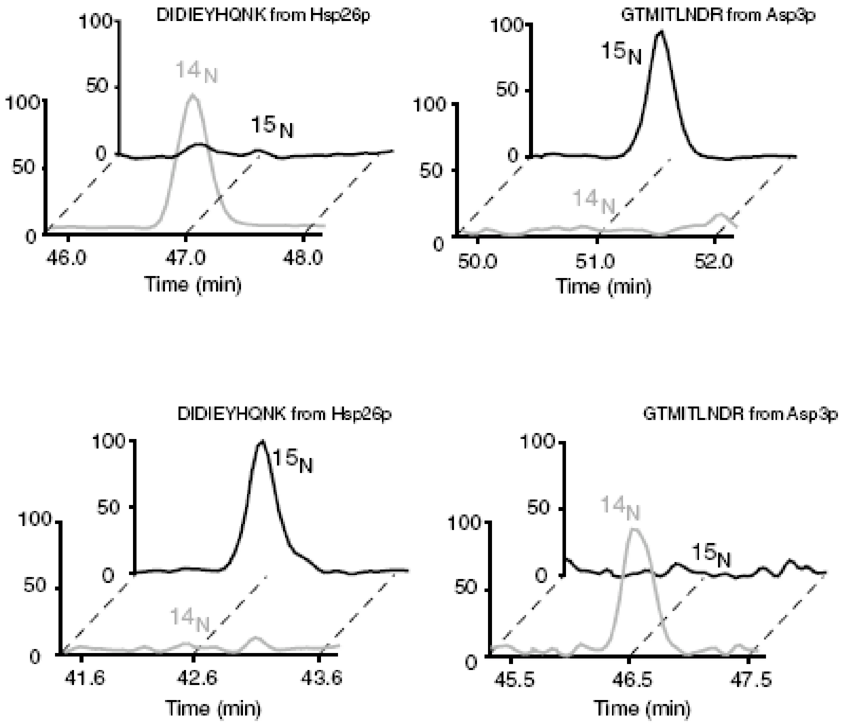


MacCoss *et al.*, *Anal Chem* **75** (2003) 6912-6921

ON/OFF peptides

"forward"
labelling

reverse
labelling



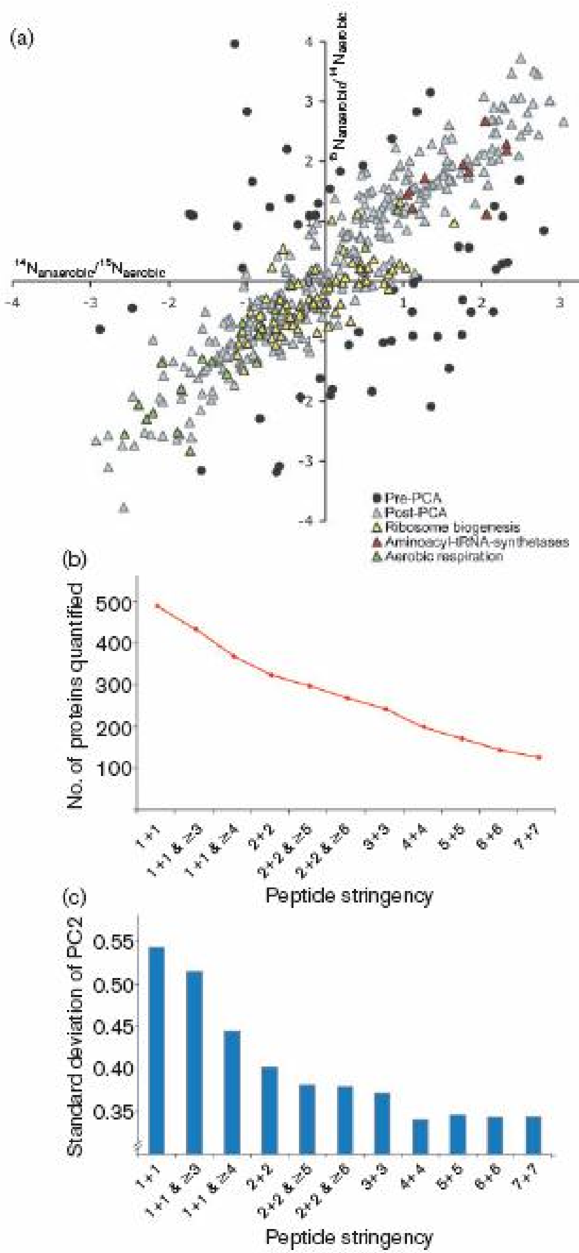
Kolkmann *et al.*, *Mol Syst Biol* **2** (2006) 2006.0026.

Generation of a robust data set

- Peptide stringency?
- PCA for inter-experimental variation (0.34)

  the STD of pc2 converges to 0.34 with increasing peptide stringency
- ...why PCA?

  ...any other suggestions?



Generation of the final data set

- # identified proteins in the triplicate analyses of the two biological replicates = 1499
- # identified proteins passed the RelEx phase = 892
- # proteins present in both biological replicates = 490
- # proteins within the 95% confidence interval of PC2 = 418
- # on/off proteins = 56
- # proteins in the final data set = 474
- # proteins with at least two-fold differential expression = 249 (137 up in anaerobiosis, 112 down in anaerobiosis)

Functional categories and subcellular locations

- Significance of over-representation of functional categories and sub-cellular locations determined using a hypergeometric test
- MIPS (Munich Information Center for Protein Sequances) functional catalogue database (FunCatDB), 28 main branches, a hierarchical, tree like structure with up to six levels of increasing specificity

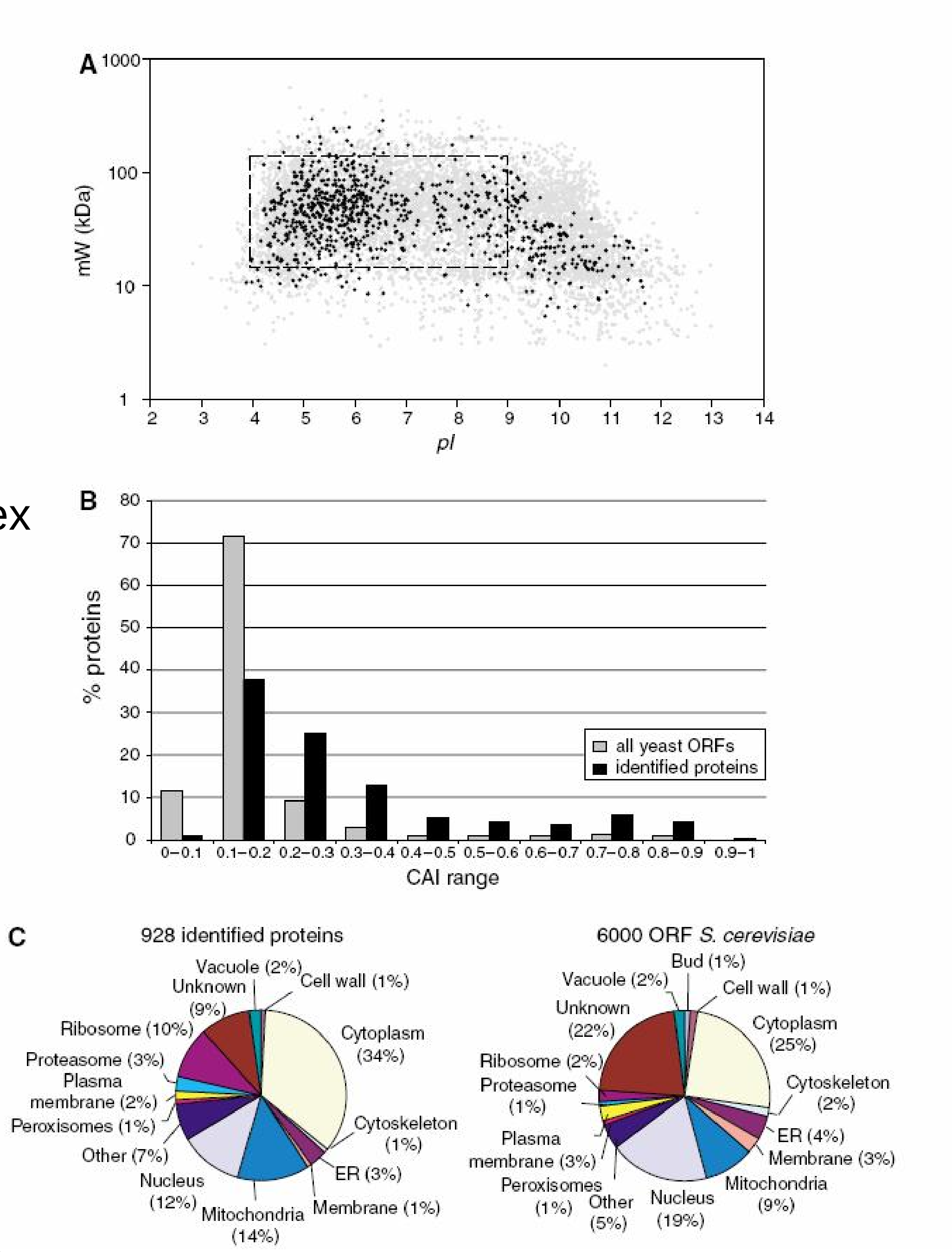
Table 1. Key functional categories of up- or downregulated proteins under anaerobiosis

MIPS category*	Protein identity	No. of proteins in cluster†	No. of proteins in genome‡	*P*-value§
Upregulated proteins				
METABOLISM (01)		92	1520	1.3×10^{-10}
Amino acid metabolism (01.01)	Gdh3p, His7p, Aro4p, Leu2p, His4p, Thr4p, Hom2p, Lys4p, Sam2p, Sah1p, Ser3p, Trp2p, Met6p, Met10p, Leu1p, Trp5p, Aro8p, Asn2p, Ade3p, Arg4p, Ser33p, Ilv3p, Cpa2p, Hom6p, Mae1p, Shm2p, Met17p, Aco1p, Ilv5p, Ilv2p, Arg1p, Ser1p, Gln1p, Tkl1p, Asn1p	35	245	7.8×10^{-5}
Purine nucleotide anabolism (01.03.01.03)	His7p, His4p, Ade5,7p, Ade6p, Ade3p, Imd2p, Ade13p, Ade17p, Ade4p, Ade2p, Ser1p	11	29	3.6×10^{-5}
C-compound and carbohydrate utilization (01.05.01)	Pyk1p, Tps1p, Adh5p, Pgi1p, Aro4p, Sam2p, Emi2p, YEL047Cp, Dld3p, Sah1p, Hxk1p, Pyc1p, Hxk2p, Ade3p, Pfk1p, Eno1p, YGR287Cp, Mal12p, Eno2p, Rhr2p, Suc2p, Tdh1p, Tdh2p, Mae1p, Pgm1p, Gpm1p, Pdc1p, Shm2p, Pdc5p, Acs2p, Aco1p, Dak1p, Pgm2p, Ilv2p, Ade17p, Pfk2p, Gpd2p, Adh1p, Fum1p, Gln1p, Tkl1p	41	388	7.9×10^{-5}
Other subcategories in METABOLISM	Imd1p, Pho88p, Cdc48p, Ssb1p, Hem13p, Rib3p, Erg1p, Rnr4p, Erg11p, Ths1p, Kar2p, Ssc1p, Grr1p, Stm1p, Yta12p, Erg2p, Faa4p, Ssb2p, Cmk2p, Hsp82p	20	ns‖	ns‖
ENERGY (02)	Pyk1p, Gdh3p, Tps1p, Adh5p, Pgi1p, Dld3p, Hxk1p, Pyc1p, Hxk2p, Ade3p, Pfk1p, Eno1p, YGR287Cp, Mal12p, Eno2p, Oye2p, Tdh1p, Aco2p, Tdh2p, Pgm1p, Gpm1p, Pdc5p, Acs2p, Aco1p, Pgm2p, Asc1p, Pfk2p, Adh1p, Hsp82p, Fum1p, Tkl1p, Rib3p, YEL047Cp, Yta12p	36	369	2.4×10^{-2}
AMINOACYL-tRNA SYNTHETASES (12.10)	Ils1p, Grs1p, Ses1p, YDR341Cp, Frs2p, Gus1p, Vas1p, Ded81p, YHR020Wp, Ths1p, Dps1p, Ala1p	12	39	2.2×10^{-7}
Downregulated proteins				
ENERGY (02)		33	369	1.7×10^{-2}
Electron transport (02.11)	Cox2p, Atp5p, Qcr7p, Rip1p, Qcr6p, Cox4p, Cox13p, Cox6p, Cyc1p, Atp7p, Sdh2p, Cox12p, Cyb2p, Cox5Ap, Cyt1p, Atp4p, Atp20p, Qcr2p	18	61	6.3×10^{-8}
Respiration (02.13)	Atp5p, Gut2p, Cyc1p, Atp7p, Cyb2p, Ald4p, Atp4p, Atp20p, Cox2p, Pet9p, Qcr7p, Rip1p, Qcr6p, Cox4p, Cox13p, Cox6p, Sdh2p, Cox12p, Cox5Ap, Cyt1p, Qcr2p	21	138	1.0×10^{-6}
Aerobic respiration (02.13.03)	Cox2p, Pet9p, Qcr7p, Rip1p, Qcr6p, Cox4p, Cox13p, Cox6p, Sdh2p, Cox12p, Cox5Ap, Cyt1p, Qcr2p	13	77	8.3×10^{-5}
Other subcategories in ENERGY	Acs1p, Mdh3p, Kgd2p, Agx1p, Pox1p, Idp2p, Adh2p, Gre2p, Lsc1p, Fdh2p, YPL276Wp, Icl2p	12	ns‖	ns‖

simulated 2D gels

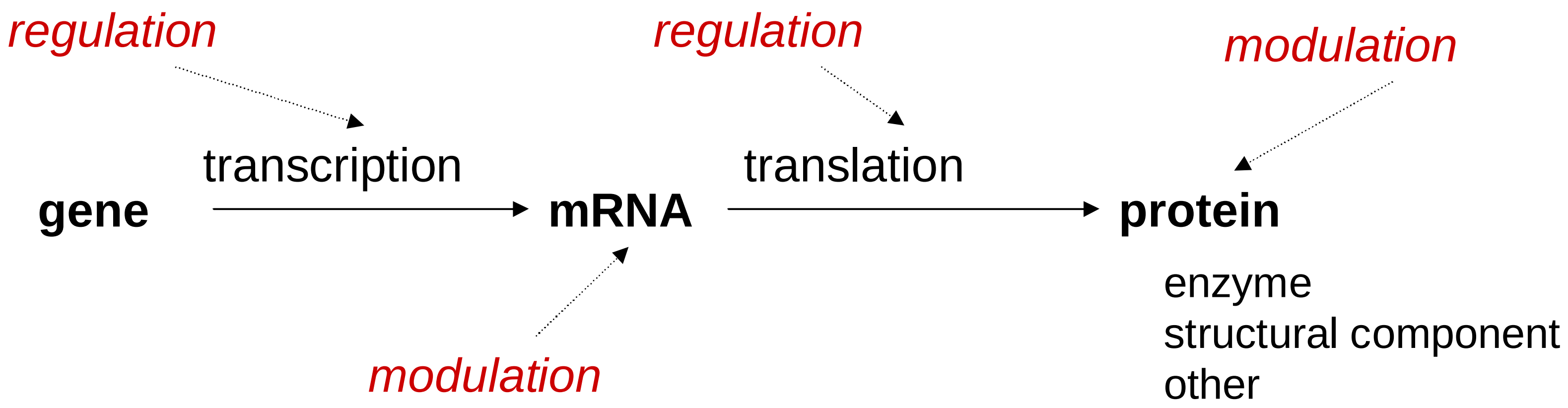
CAI = codon adaptation index

Subcellular localisation



Kolkmann *et al.*, *Mol Syst Biol* **2** (2006) 2006.0026.

Regulatory level?



regulation

regulation

modulation

transcription

translation

gene

mRNA

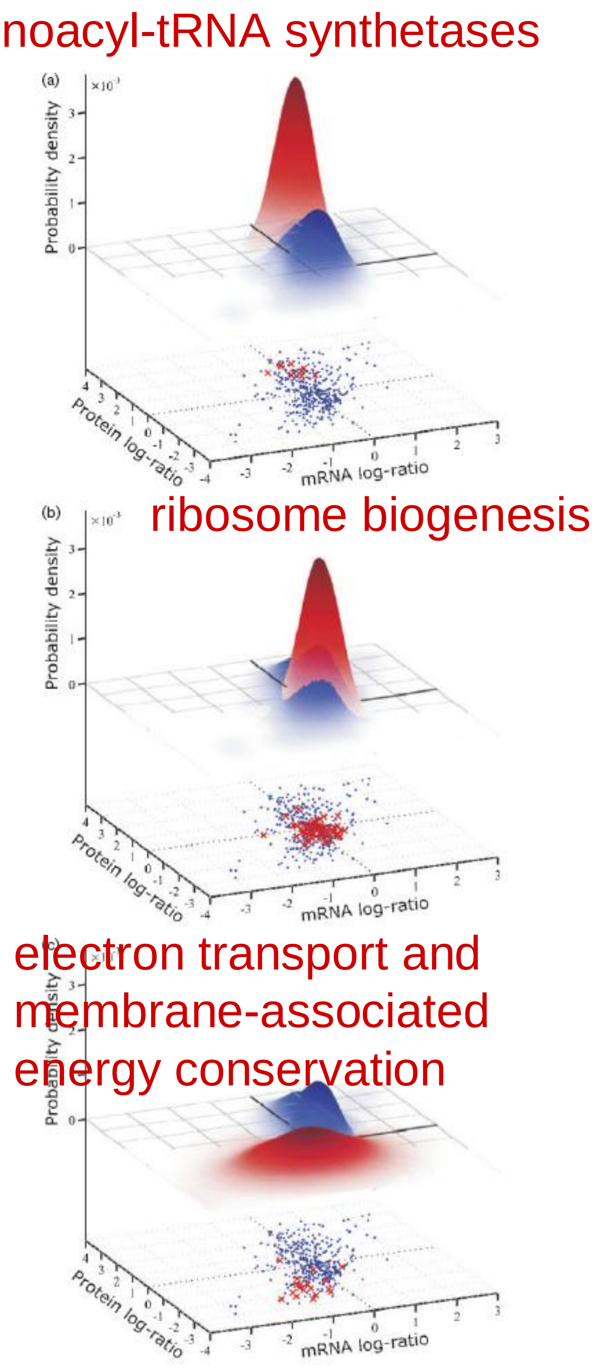
protein

enzyme
structural component
other

modulation

Probability density test

- Transcript data from:
  Tai *et al.*, *J Biol Chem* **280** (2005) 437-447.
- Were the proteins of a specific MIPS category enriched in a specific part of the data space covering the protein vs the mRNA ratios?
- H_0: the data points corresponding to a particular MIPS category are randomly sampled from all proteins
- To test H_0 the PDF of the complete distribution was compared to the PDF of a particular MIPS category
- PDFs were estimated and evaluated at a grid of 2500 co-ordinates
- RMS value of the difference between the complete set and a particular MIPS category was computed
- Permutation tests for significance to the possible rejection of H_0

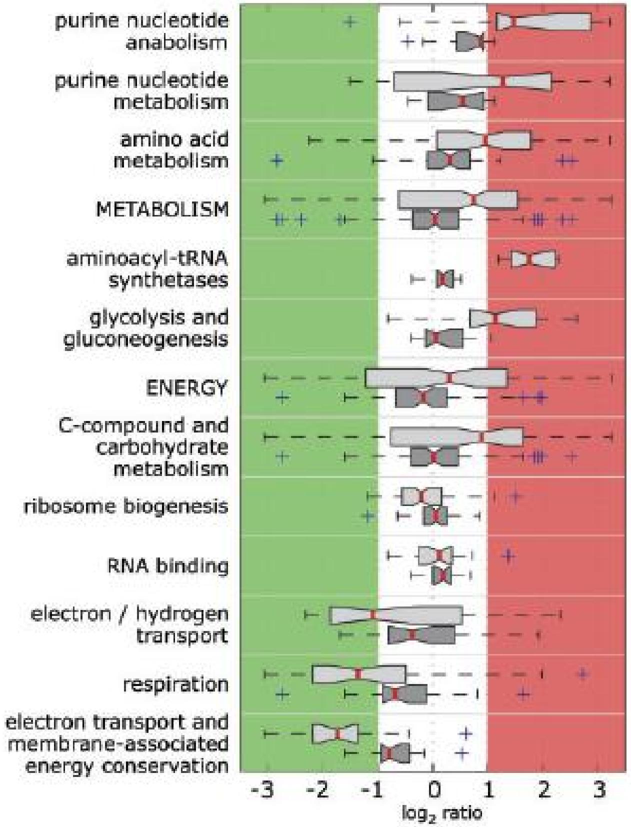

aminoacyl-tRNA synthetases

ribosome biogenesis

electron transport and membrane-associated energy conservation

Regulatory level box plot

- Significant functional categories based on the probability density test



Conclusions

- Quantitative protein data is required for studying cell regulatory functions
- Quantitative data in relative form
- The main data processing steps:

  i) assignment of the fragment ion spectra to peptide sequences
  ii) inference of the proteins represented by the identified peptides
  iii) determination of the abundancies of the proteins

- Data processing is still in stage of development

  data processing includes manual steps

  assessment of the quality of the data?

- Integration of omics-data