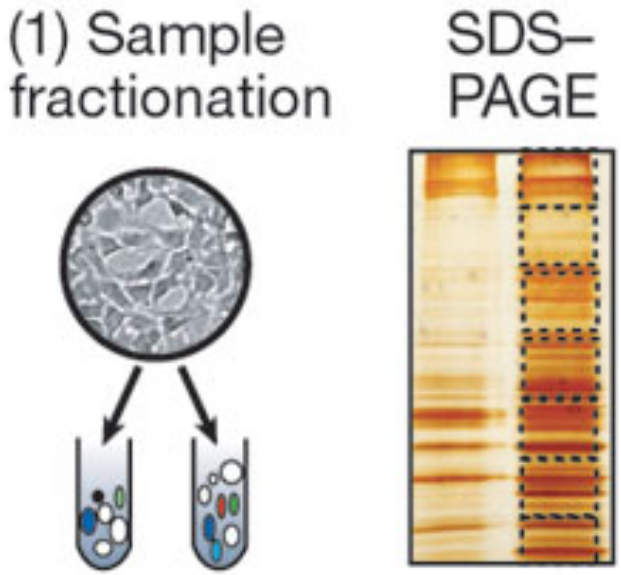
V4 – MS proteomics – data imputation

- How does MS proteomics work?
- What is the role of bioinformatics in MS proteomics ?
  - Peptide mass fingerprinting
  - Significance analysis
  - GO annotations
- Applications of MS:
  - TAP-MS
  - Phosphoproteome
  - Cell-cycle oscillatory proteins
- **Data imputation** for MS data

Noble prize in chemistry 2002
John B. Fenn Koichi Tanaka
"for their development of soft desorption ionisation methods for mass spectrometric analyses of biological macromolecules"

www.nobelprize.org

Proteomics workflow



The typical proteomics experiment consists of 5 stages.

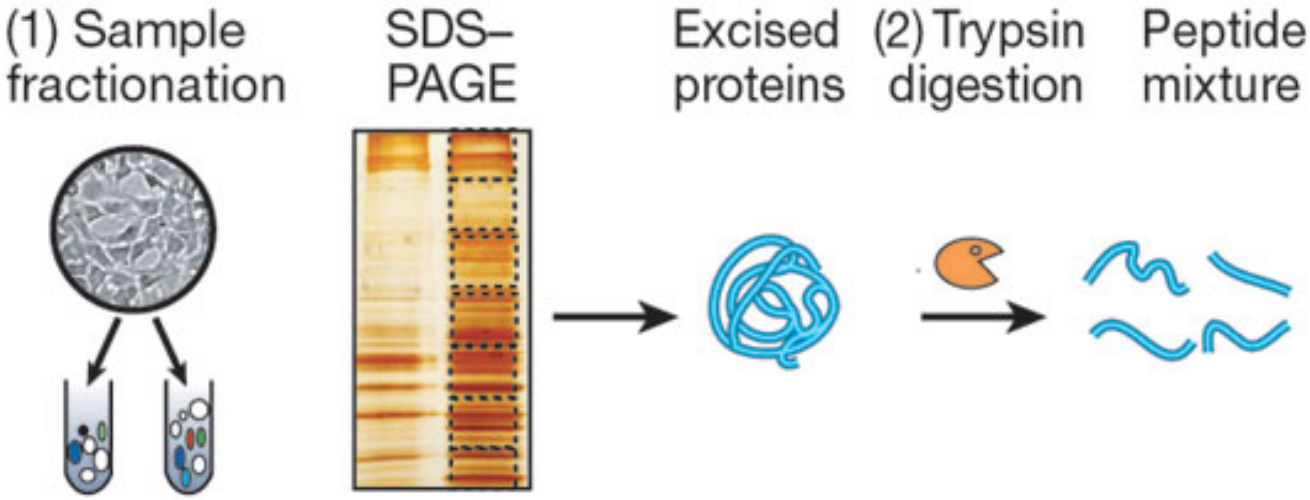
In stage 1, the proteins to be analyzed are **isolated** from cell lysate or tissues by biochemical fractionation or affinity selection.

This often includes a final step of one-dimensional gel electrophoresis, and defines the 'sub-proteome' to be analysed.

MS of whole proteins is less sensitive than **peptide MS** and the mass of the intact protein by itself is insufficient for identification.

Aebersold, Mann
Nature 422, 198-207(2003)

Proteomics workflow



Therefore, proteins are **degraded enzymatically** to peptides in stage 2, usually by trypsin, leading to peptides with C-terminally protonated amino acids (K/R), providing an advantage in subsequent peptide sequencing.

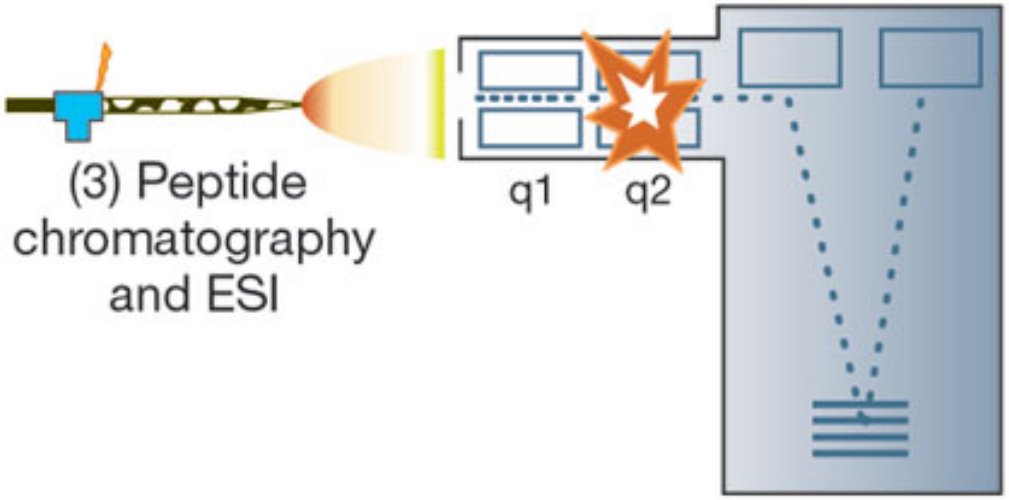
Table 1. Distrubution of peptide fragment length from 20,639 proteins

Enzyme/reagent	Residues cleaved	Total fragments	Avg. fragment length
Trypsin	K/R	662,981	8
Lys-C	K	359,140	16
Asp-N	D	321,655	18
CNBr	M	150,605	38
Hydroxylamine	N-G	36,643	152
Dilute acid	D-P	35,574	166

Aebersold, Mann
Nature 422, 198-207(2003)

Henzel et al. J Am Soc Mass Spectrom
14, 931–942 (2003)

Proteomics workflow



In stage 3, the peptides are **separated** by one or more steps of high-pressure liquid chromatography in very fine capillaries.

Then, they are eluted e.g. into an electrospray ion source where they are nebulized in small, highly charged droplets.

After evaporation, multiply protonated peptides enter the mass spectrometer.

Aebersold, Mann
Nature 422, 198-207(2003)

Mass spectrometer

A mass spectrometer consists of an **ion source**, a **mass analyser** that measures the **mass-to-charge ratio** (*m*/*z*) of the ionized analytes, and a **detector** that registers the number of ions at each *m*/*z* value.

Electrospray ionization (ESI) and **matrix-assisted laser desorption/ionization** (MALDI) are the two techniques most commonly used to volatize and ionize the proteins or peptides for mass MS analysis.

ESI ionizes the analytes out of a solution and is therefore readily coupled to liquid-based (e.g. chromatographic and electrophoretic) separation tools.

MALDI sublimates and ionizes the samples out of a dry, crystalline matrix via laser pulses. MALDI-MS is normally used to analyse relatively simple peptide mixtures, whereas integrated liquid-chromatography ESI-MS systems (LC-MS) are preferred for the analysis of complex samples

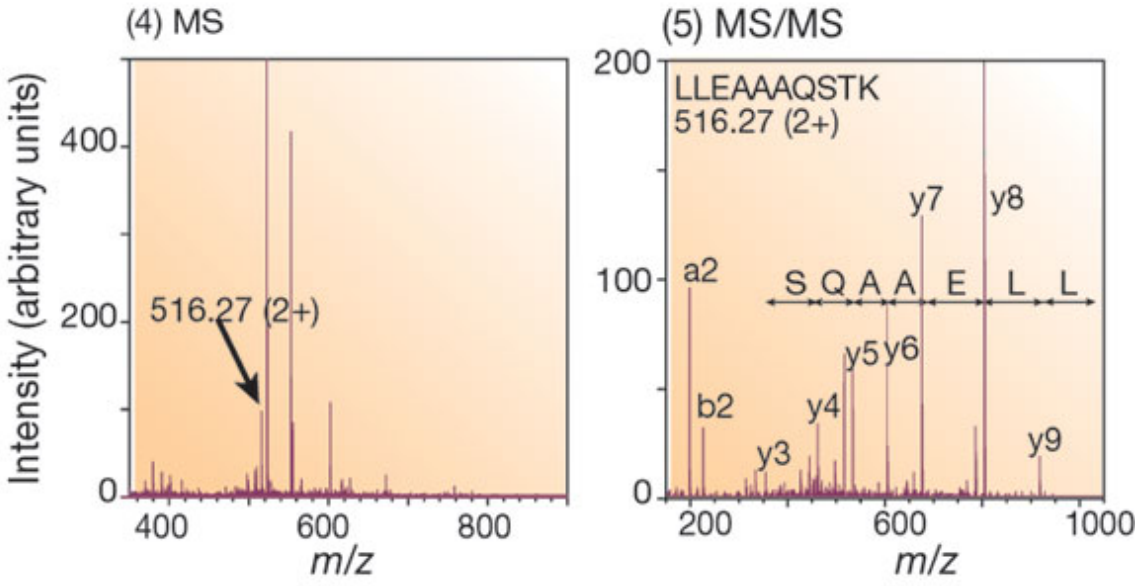
Aebersold, Mann
Nature 422, 198-207(2003)

Proteomics workflow

In stage 4, a mass spectrum of the peptides eluting at this time point is taken.

Mass peak ≡ sequence **composition** of a peptide.

The computer the generates a prioritized list of the peptides for a second fragmentation.



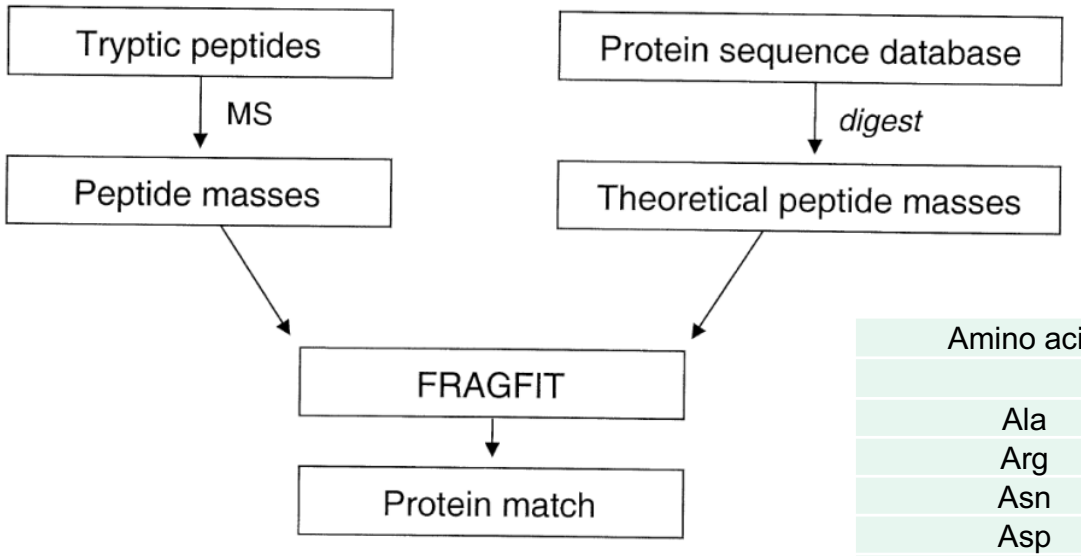
In stage 5, a series of tandem mass spectrometric or 'MS/MS' experiments is performed to determine the sequence of a peptide (here, the peak *m* = 516.27 Da).

The MS and MS/MS spectra are matched against protein sequence databases ("**peptide mass fingerprinting**").

The outcome of the experiment is the identity of the peptides and therefore the proteins making up the purified protein population.

Aebersold, Mann
Nature 422, 198-207(2003)

Peptide mass fingerprinting



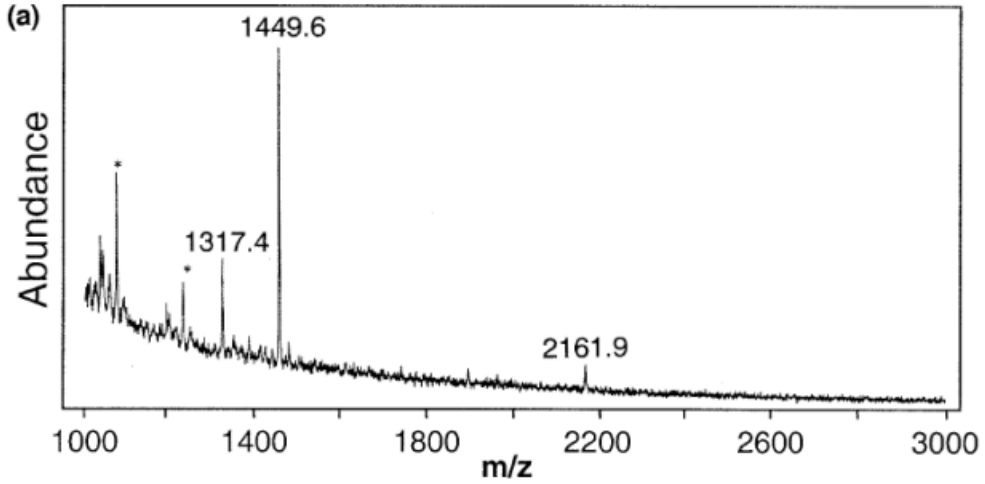
$$m_{peptide} = \sum_{i \in amino\ acids\ 1 \ldots n} m_i$$

Amino acid	Mono-Isotopic mass [Da]	Average mass [Da]
Ala	71.037114	71.0779
Arg	156.101111	156.1857
Asn	114.042927	114.1026
Asp	115.026943	115.0874
Cys	103.009185	103.1429
Glu	129.042593	129.114
Gln	128.058578	128.1292
Gly	57.021464	57.0513
His	137.058912	137.1393
Ile	113.084064	113.1576
Leu	113.084064	113.1576
Lys	128.094963	128.1723
Met	131.040485	131.1961
Phe	147.068414	147.1739
Pro	97.052764	97.1152
Ser	87.032028	87.0773
Thr	101.047679	101.1039
Trp	186.079313	186.2099
Tyr	163.06332	163.1733
Val	99.068414	99.1311

The masses of peptides from a database are compared with experimentally determined masses using a software.

Henzel et al. J Am Soc Mass Spectrom 14, 931–942 (2003);
www.matrixscience.com

Peptide mass fingerprinting



(b)
```
enzyme: Asp-N (N-side of Asp)
Mass of MH+: 1317.400 1449.600 2161.900 (tol: 1.000)
LZCH Lysozyme c (EC 3.2.1.17) precursor - Chicken
         2162.444          84: DGRTPGSRNLCNIPCSALLSS
         1449.706         105: DITASVNCAKIVS
         1317.552         137: DVQAWIRGCRL
```

Mass [Da]

Starting position

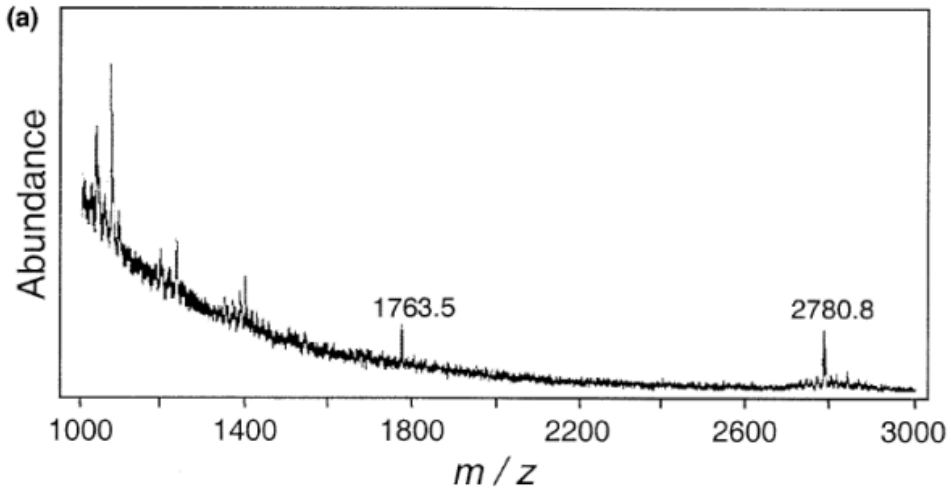
Peptide fragment

(**a**) FAB ("fast atom bombardment" = old technique) spectrum of a 250 pmol tryptic digest of Asp-N digest of **lysozyme**.

(**b**) FRAGFIT output page showing a match with chicken egg white lysozyme obtained using the masses from the MS spectrum.

Henzel et al. J Am Soc Mass Spectrom 14, 931–942 (2003)

Peptide mass fingerprinting



(b)

```
enzyme: CNBr (C-side of Met)
Mass of MH+: 1763.500 2780.800 (tol: 0.600)
CCHO Cytochrome C - Horse
          1764.031          66: EYLENPKKYIPGTKM
          2781.268          81: IFAGIKKKTEREDLIAYLKKATNE
CCHOD Cytochrome C - Donkey and common zebra
                     (tentative sequences)
          1764.031          66: EYLENPKKYIPGTKM
          2781.268          81: IFAGIKKKTEREDLIAYLKKATNE
```

(**a**) FAB spectrum of a 500 pmol CNBr cleavage of horse heart **cytochrome *c***.

(**b**) FRAGFIT output page showing a match with cytochrome *c* obtained using the masses from the FAB spectrum.

The output includes all proteins that match the mass list.

The 2 masses observed were sufficient to identify the protein as cytochrome c and permitted the identification of the species.

At the time this search was performed, the database contained nearly 100 different species of cytochrome *c*

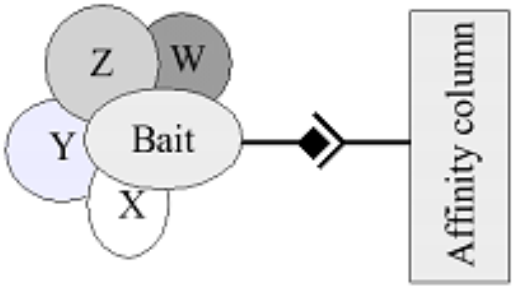
Henzel et al. J Am Soc Mass Spectrom 14, 931–942 (2003)

Application: Detect protein-protein interactions: Tandem affinity purification (also „pull-down“)

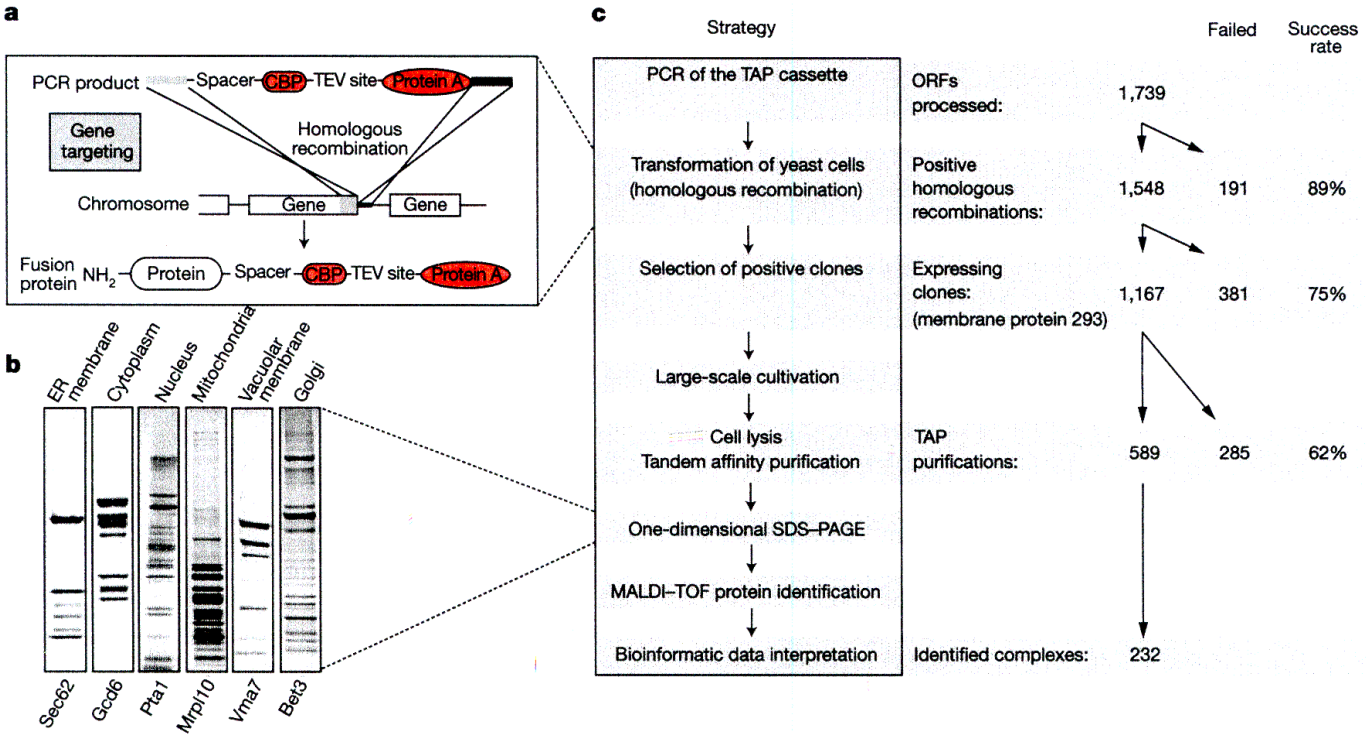
In **affinity purification**, a protein of interest (bait) is tagged with a molecular label (dark route in the middle of the figure) to allow easy purification.

The tagged protein is then co-purified together with its interacting partners (W–Z).

This strategy can be applied on a genome scale (as Y2H).



Identify proteins by mass spectro-metry (MALDI-TOF).

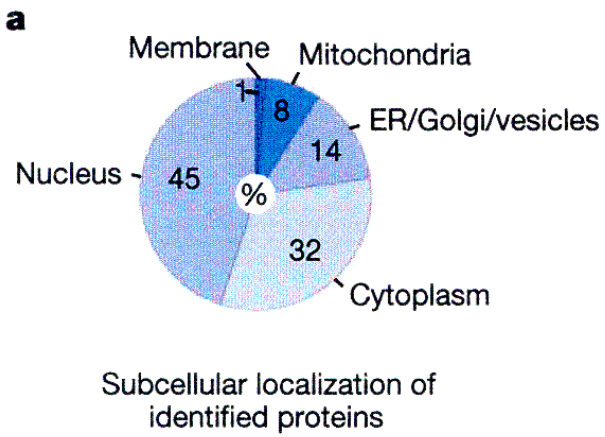


Gavin *et al. Nature* 415, 141 (2002)

TAP analysis of yeast PP complexes

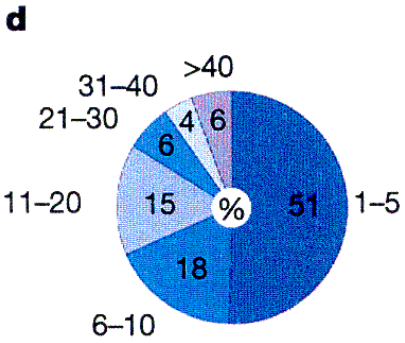
Identify proteins by scanning yeast protein database for protein composed of fragments of suitable mass.

(a) lists the identified proteins according to their localization
-> no apparent bias for one compartment, but very few membrane proteins (should be ca. 25%)



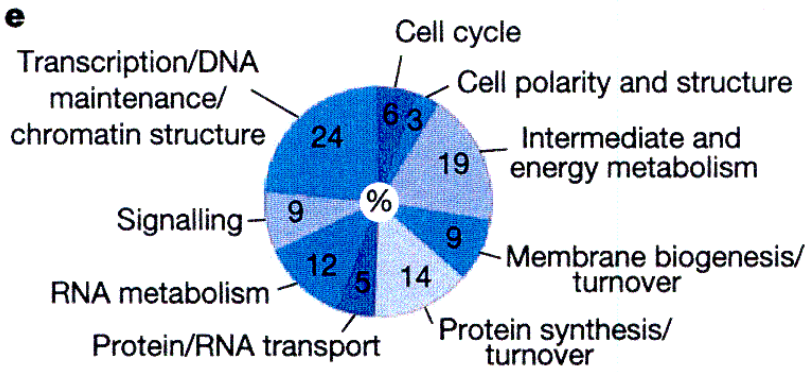
Subcellular localization of identified proteins

(d) lists the number of proteins per complex
-> half of all PP complexes have 1-5 members, the other half is larger



Number of proteins per complex

(e) Complexes are involved in practically all cellular processes



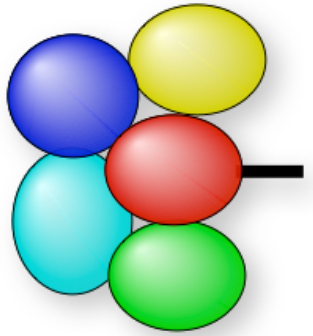
Distribution of complexes according to function

Gavin *et al. Nature* 415, 141 (2002)

Pros and Cons of TAP-MS

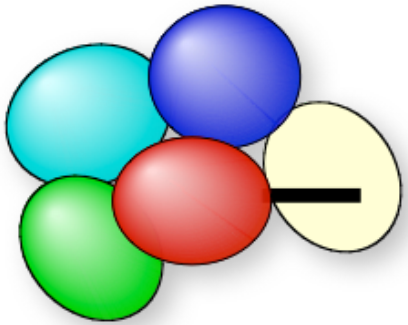
Advantages:

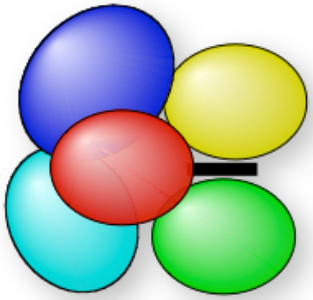
- **quantitative** determination of complex partners ***in vivo*** without prior knowledge
- simple method, high yield, high throughput



Difficulties:

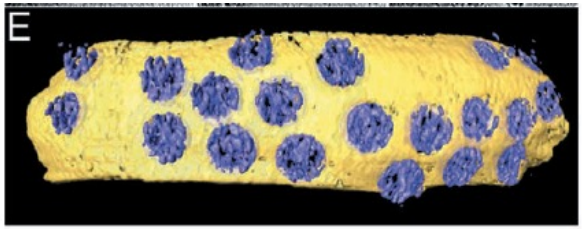
- tag may **prevent** binding of the interaction partners
- tag may change (relative) **expression** levels
- tag may be **buried** between interaction partners
  → no binding to beads



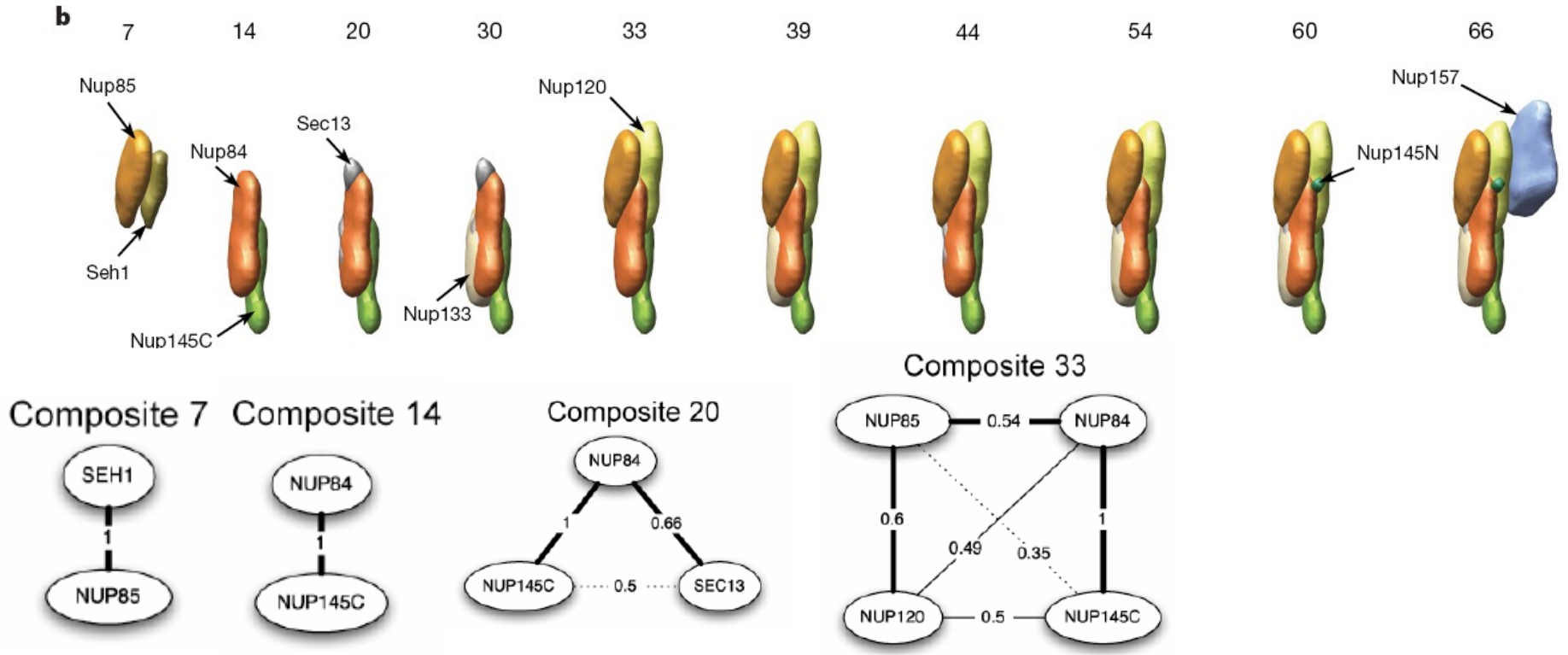


Protein interactions in nuclear pore complex

Figure (right) shows 20 NPCs (blue) in a slice of a nucleus.
Aim: identify individual PPIs in Nuclear Pore Complex.

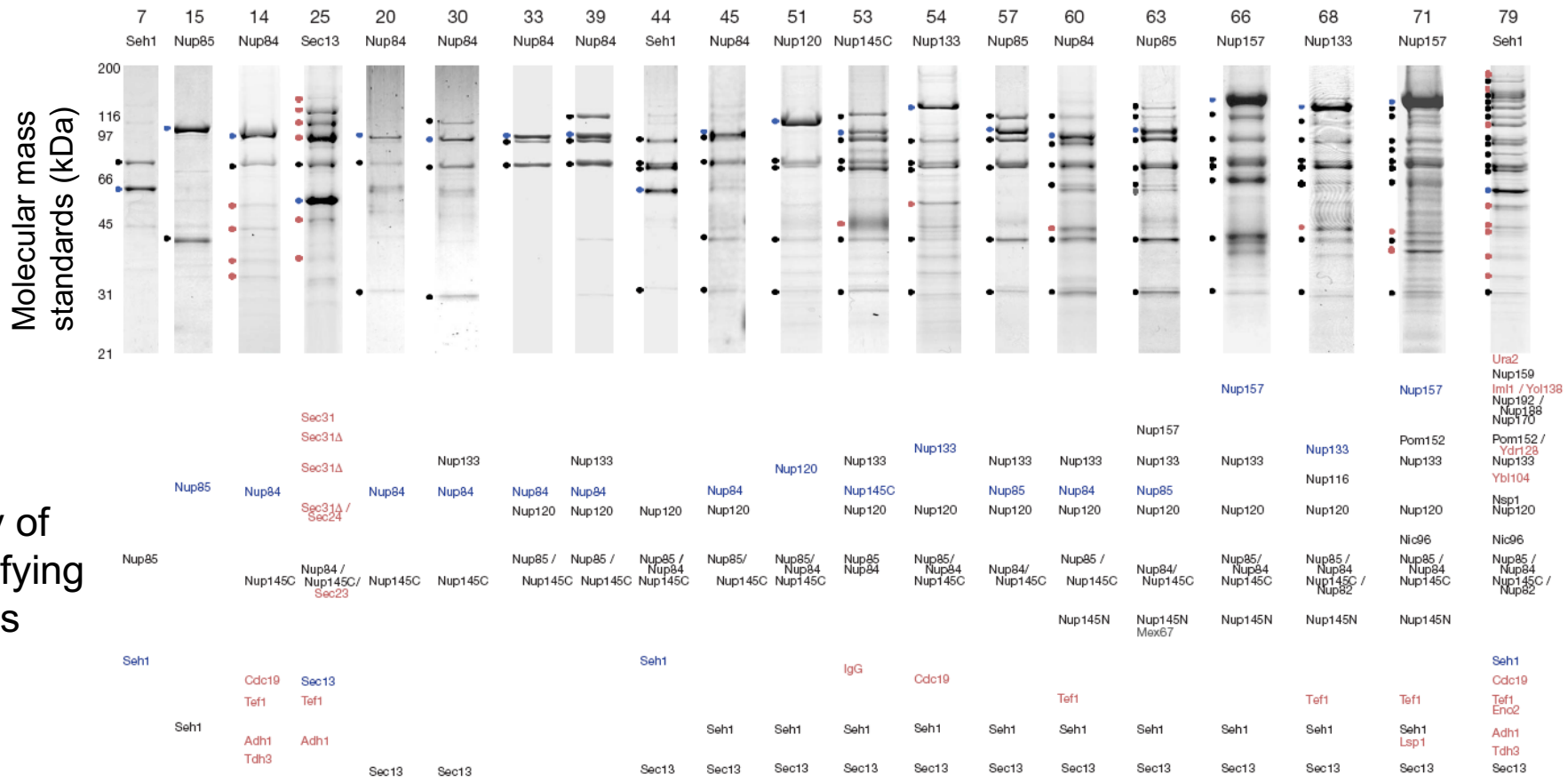


Below : mutual arrangement of Nup84-complex-associated proteins as visualized by their localization volumes in the final NPC structure.
Nup84 protein shown in light brown.



SDS + MS:Composites involving Nup84

a above lanes: name of ProteinA-tagged protein and identification number for composite



Blue: PrA-tagged proteins,
Black: co-purifying nucleoporins,
Grey: NPC-associated proteins,
Red: and other proteins (e.g. contaminants)

Affinity-purified PrA-tagged proteins and interacting proteins were resolved by **SDS–PAGE** and visualized with Coomassie blue. The bands marked by filled circles at the left of the gel lanes were identified by **mass spectrometry** (cut out band from the gel and use as input for MS).

Application: Protein phosphorylation during cell cycle

Protein **phosphorylation** and **dephosphorylation** are highly controlled biochemical processes that respond to various intracellular and extracellular stimuli.

Phosphorylation status modulates protein activity,
- influencing the tertiary and quaternary **structure** of a protein,
- controlling **subcellular distribution**, and
- regulating **interactions** with other proteins.

Regulatory protein phosphorylation is a **transient** modification that is often of low occupancy or “stoichiometry”

This means that only a fraction of a particular protein may be phosphorylated on a given site at any particular time, and that occurs on regulatory proteins of low abundance, such as protein kinases and transcription factors.

Olsen Science Signaling 3 (2010)

Cell Cycle and the Phosphoproteome

CELL CYCLE

Quantitative Phosphoproteomics Reveals Widespread Full Phosphorylation Site Occupancy During Mitosis

Jesper V. Olsen,[1,2*] Michiel Vermeulen,[1,3*] Anna Santamaria,[4*] Chanchal Kumar,[1,5*]
Martin L. Miller,[2,6] Lars J. Jensen,[2] Florian Gnad,[1] Jürgen Cox,[1] Thomas S. Jensen,[7]
Erich A. Nigg,[4] Søren Brunak,[2,7] Matthias Mann[1,2†]

(Published 12 January 2010; Volume 3 Issue 104 ra3)

www.SCIENCESIGNALING.org 12 January 2010 Vol 3 Issue 104 ra3

Aim: Analyze all proteins that are modified by phosphorylation during different stages of the cell cycle of human HeLa cells.

Ion-exchange chromatography + HPLC + MS + sequencing led to the identification of 6695 proteins.
From this 6027 quantitative cell cycle profiles were obtained.

A total of 24,714 phosphorylation events were identified.
20,443 of them were assigned to a specific residue with high confidence.

Finding: about **70%** of all proteins get phosphorylated.

Review: protein quantification by SILAC

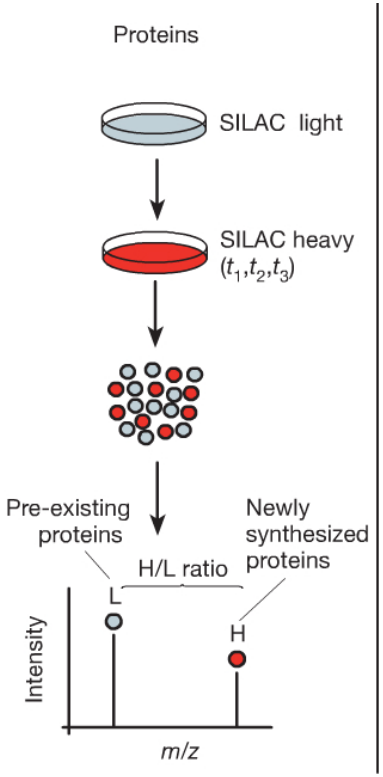
ARTICLE

doi:10.1038/nature10098

Global quantification of mammalian gene expression control

Björn Schwanhäusser[1], Dorothea Busse[1], Na Li[1], Gunnar Dittmar[1], Johannes Schuchhardt[2], Jana Wolf[1], Wei Chen[1] & Matthias Selbach[1]

SILAC: „**s**table **i**sotope **l**abelling by **a**mino acids in **c**ell culture“ means that cells are cultivated in a medium containing heavy stable-isotope versions of essential amino acids.



When non-labelled (i.e. light) cells are transferred to heavy SILAC growth medium, newly synthesized proteins incorporate the heavy label while pre-existing proteins remain in the light form.

Quantification protein turnover and levels.
Mouse fibroblasts are transferred to medium with heavy amino acids (SILAC)

.

Protein turnover is quantified by mass spectrometry and next-generation sequencing, respectively.

Schwanhäuser et al. Nature 473, 337 (2011)

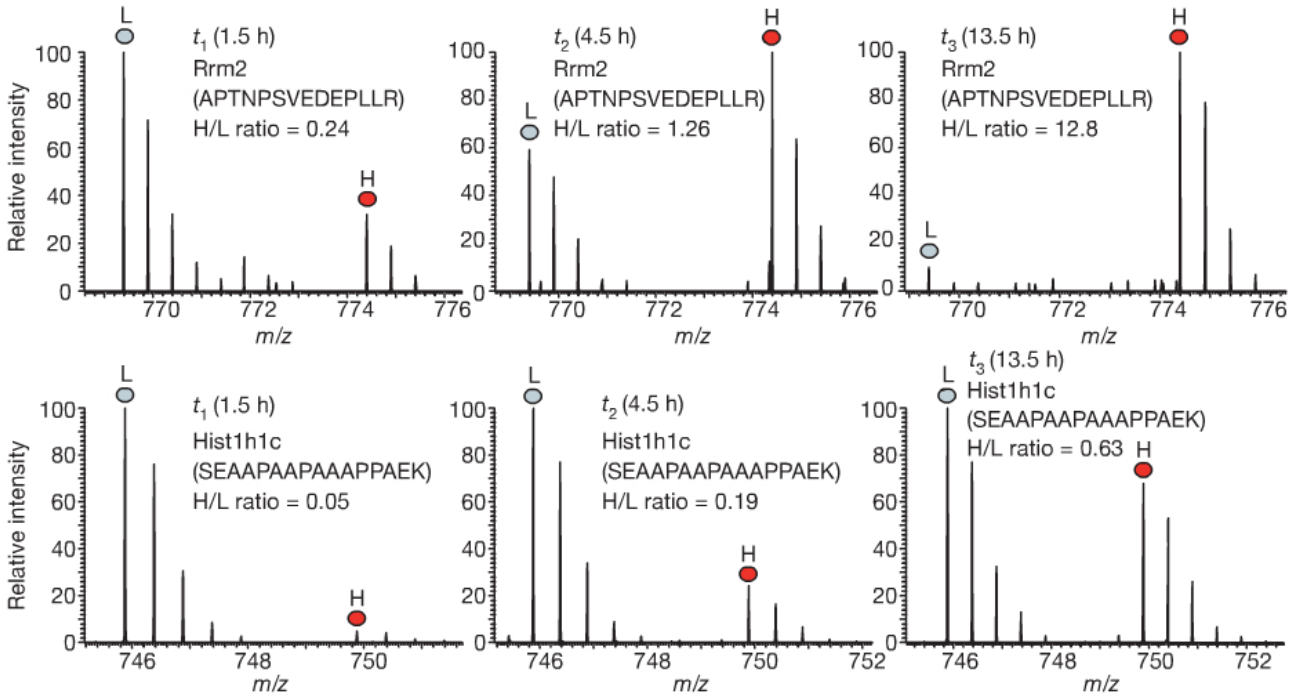
Rates of protein translation

Mass spectra of peptides for two proteins.

Top: **high-turnover protein**
Bottom: **low-turnover protein**.

Over time, the heavy to light (H/L) ratios increase.

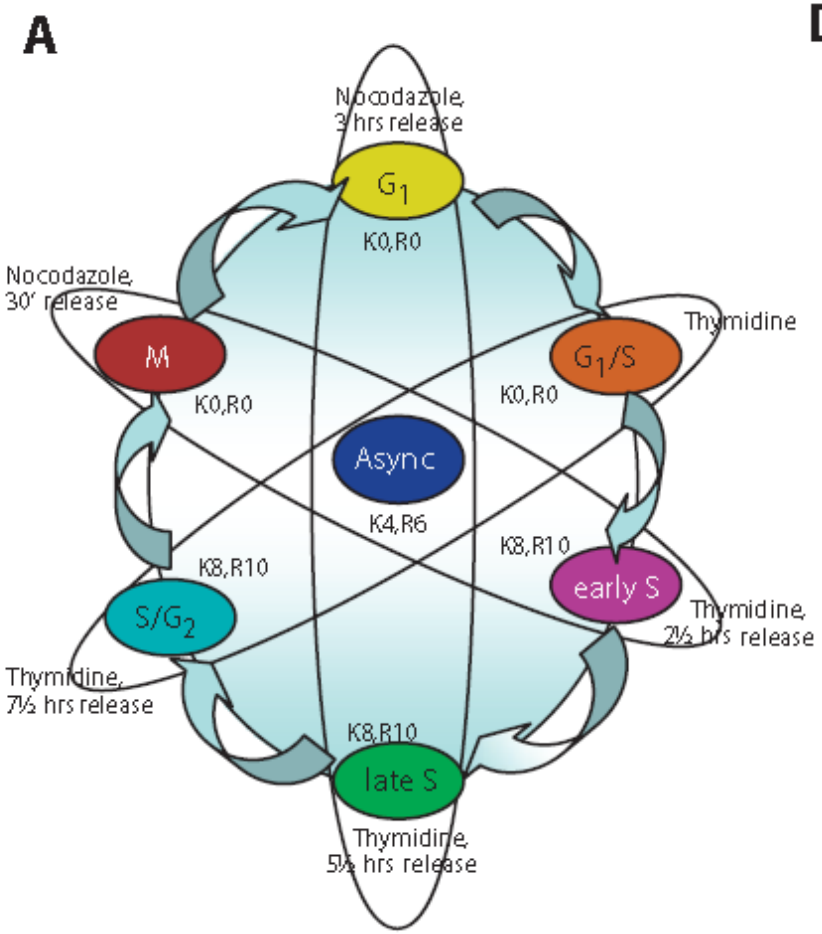
H-concentration of high-turnover protein saturates.
That of low-turnover protein still increases.



This example was introduced to illustrate the principles of SILAC and mass spectroscopy signals (peaks).
In the Olson et al. study, the authors used H and L forms to label different stages of the cell cycle.

Schwanhäuser et al. Nature 473, 337 (2011)

Quantitative proteomic analysis



- HeLa S3 cells were SILAC-labeled with 3 different isotopic forms (light – medium –heavy) of arginine and lysine.

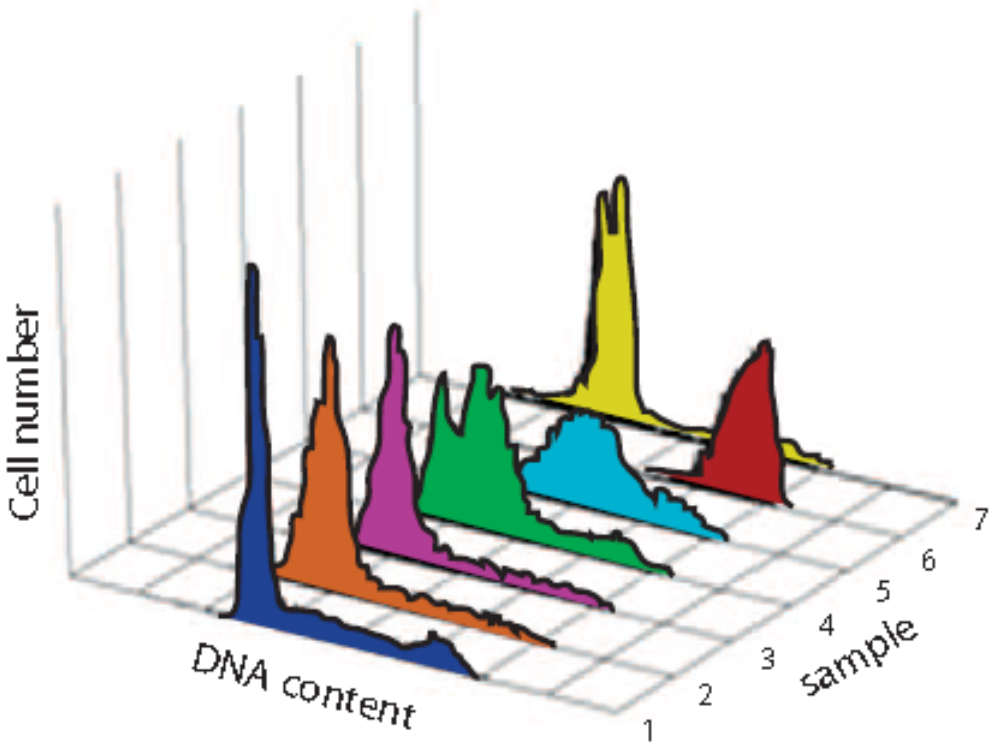
3 individual populations of heavy and light SILAC cells were synchronized with a **thymidine** block (analog of thymine, blocks entry into S phase). Cells were then collected at six different time points across the cell cycle after release from the thymidine arrest.

2 samples were collected after a **cell cycle arrest** with **nocodazole** and release. (Nocodazole interferes with polymerization of microtubules.)

Cells were lysed and mixed in equal amounts using an asynchronously growing cell population as the internal standard to allow normalization between experiments.
3 independent experiments were performed to cover six cell cycle stages.

FACS profiles of individual HeLa populations

	% Cells		
	G_1	S	G_2/M
1. Asynchronous	64	27	9
2. Thymidine block	50	46	4
3. Thymidine block + release 2½ h	36	60	4
4. Thymidine block + release 5½ h	23	70	7
5. Thymidine block + release 7½ h	15	70	15
6. Nocodazole block + release ½ h	1	11	88
7. Nocodazole block + release 3 h	82	12	6



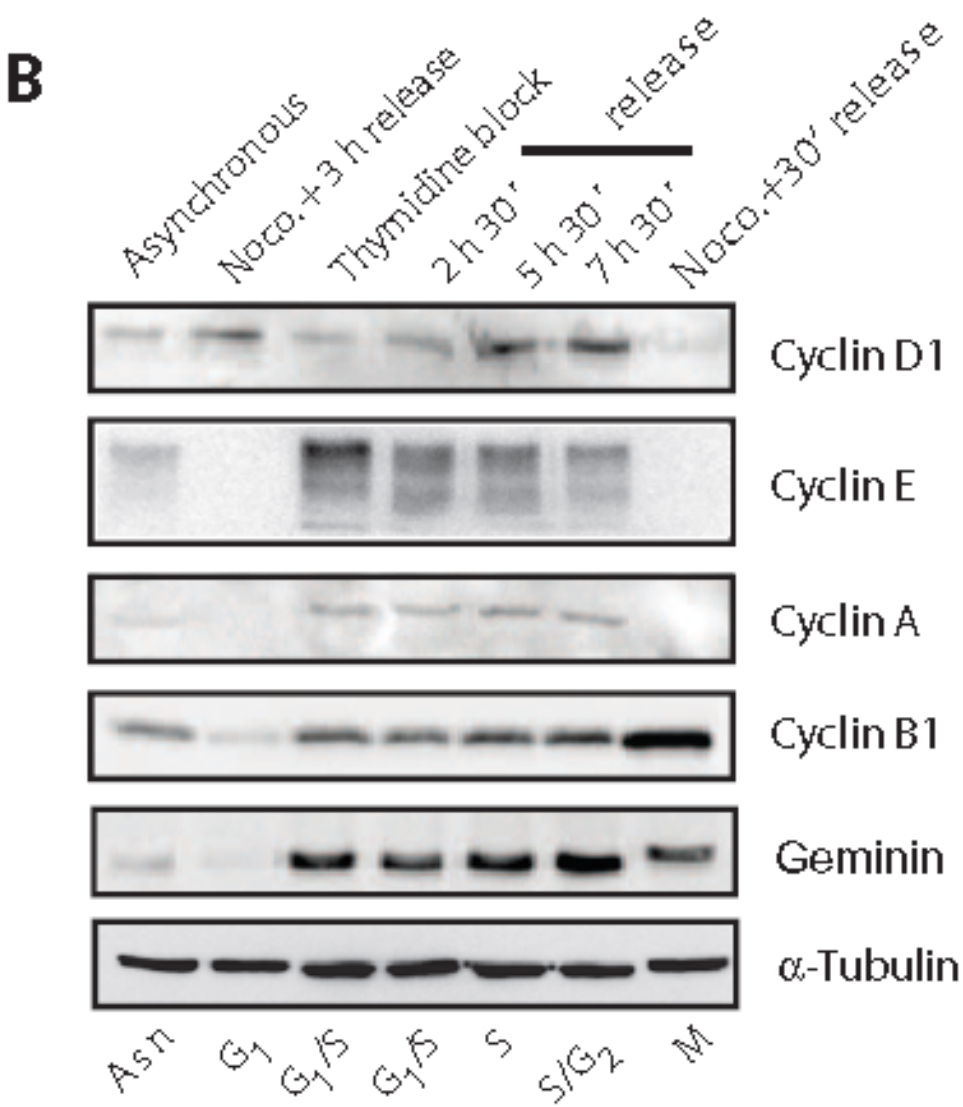
Cells were fixed and collected by centrifugation.

Then the DNA content of the cells was determined with propidium iodide.

This is the basis for classifying the state along the cell cycle.

Olsen Science Signaling 3 (2010)

Quantification of cell cycle markers

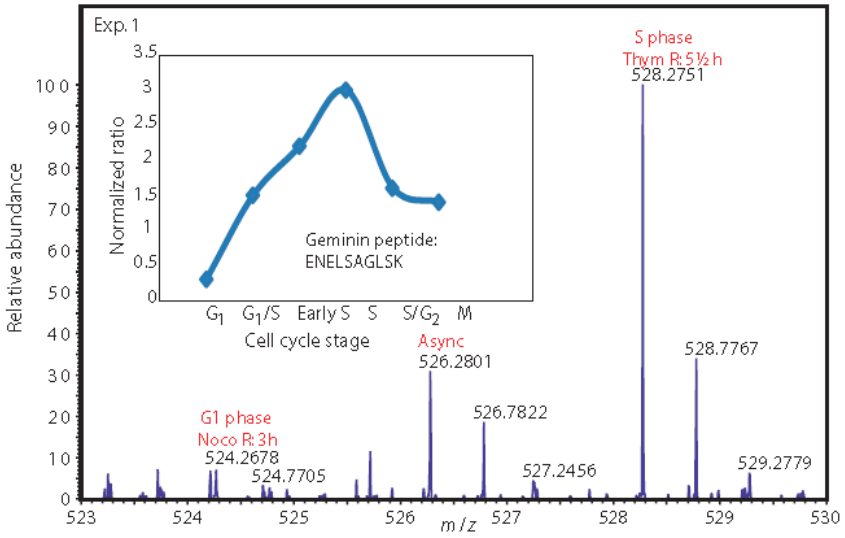


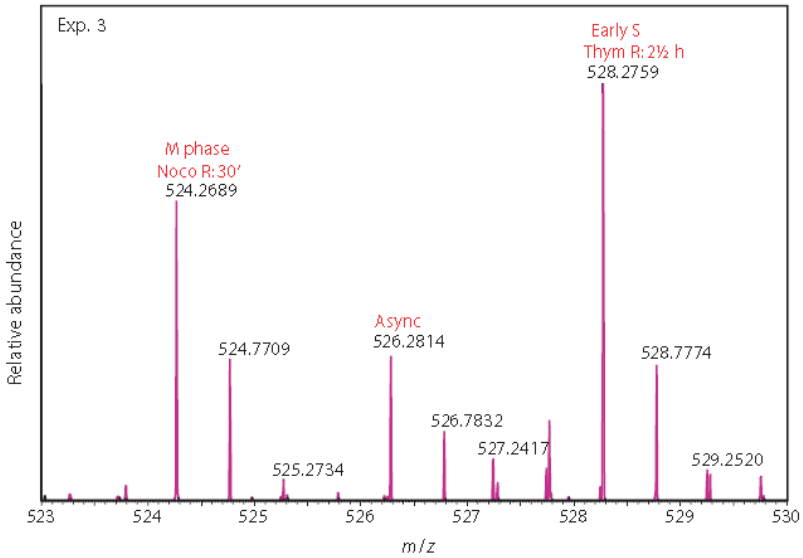
Immunoblot analysis of known cell cycle marker proteins in the different cell populations.

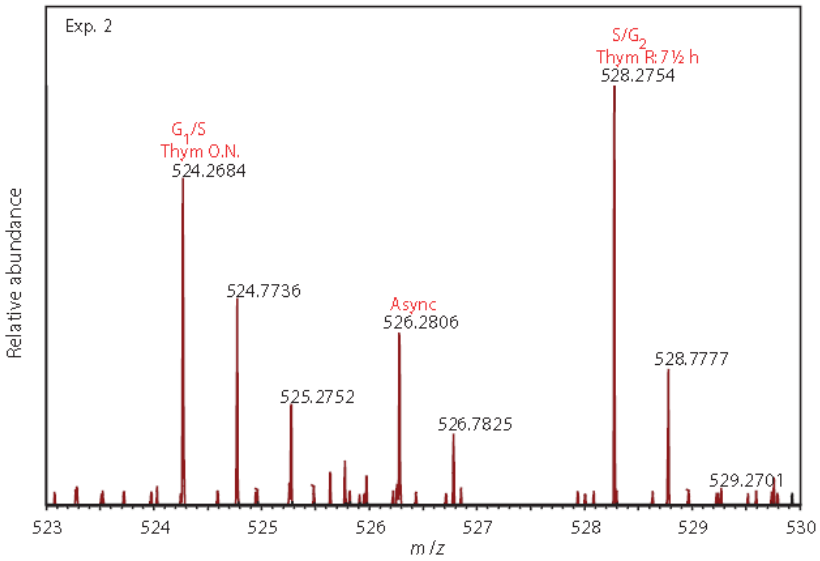
The abundance of a fifth of the proteome changed by at least fourfold throughout the cell cycle (difference between lowest and highest abundance).

Because a **fourfold change** also best accounted for the dynamics of already described cell cycle components, this ratio was used as a threshold for subsequent analysis.

Monitoring of protein abundance by MS







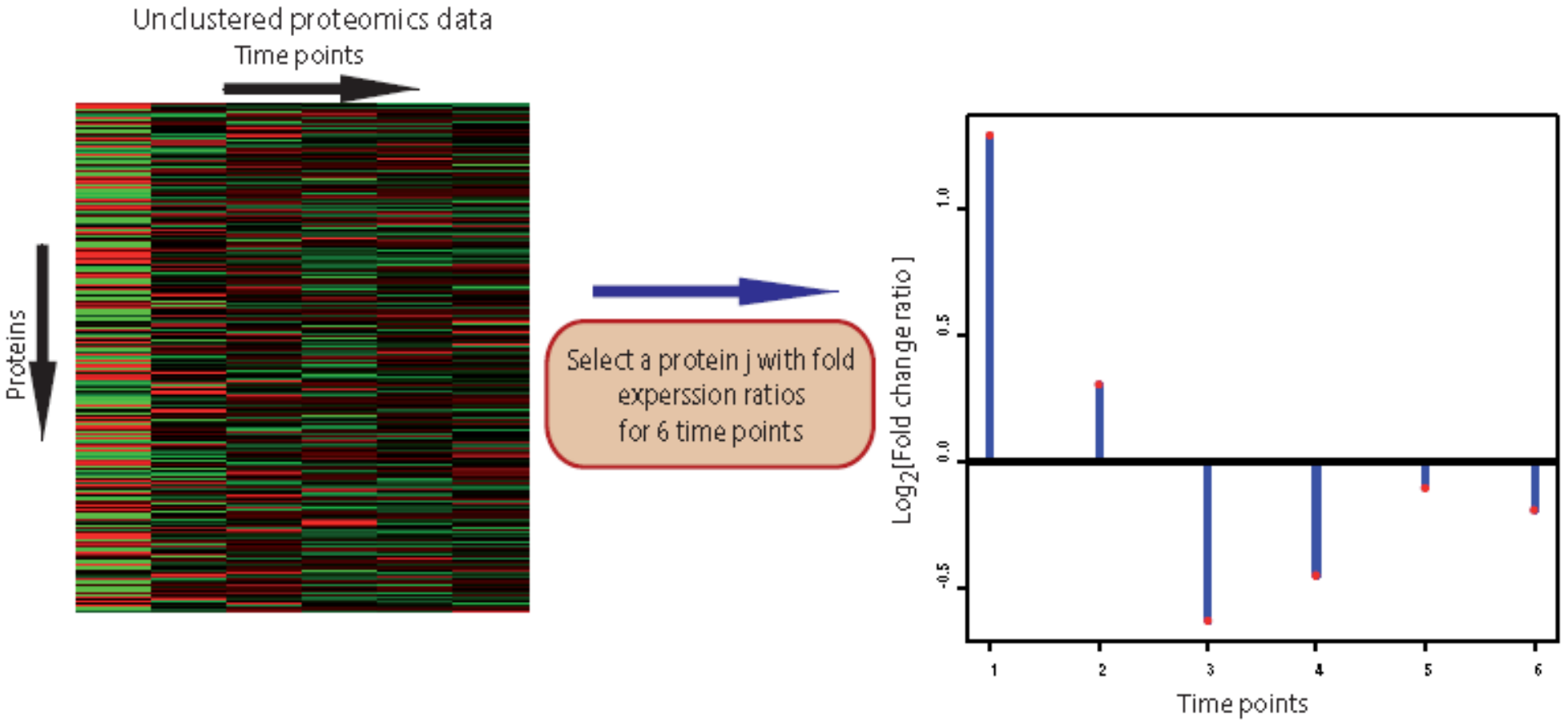
Representative MS data showing how the abundance of the proteins was monitored in three experiments (Exp. 1, Exp. 2, Exp. 3) to obtain information from the 6 stages of the cell cycle.

The data show the MS analysis of a tryptic SILAC peptide triplet derived from the cell cycle marker protein **Geminin**.

Relative peptide abundance changes were normalized to the medium SILAC peptide derived from the asynchronously grown cells in all three experiments. The inset shows the combined six-time profile of Geminin over the cell cycle.

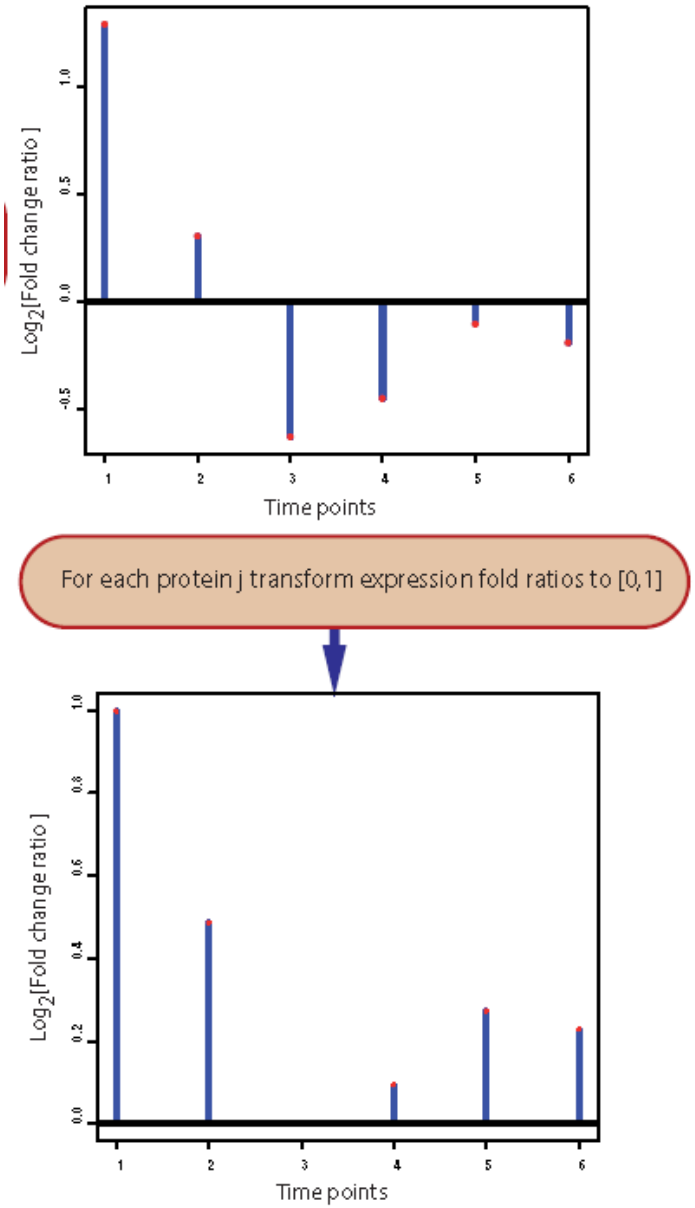
Olsen Science Signaling 3 (2010)

Bioinformatics Workflow (1)



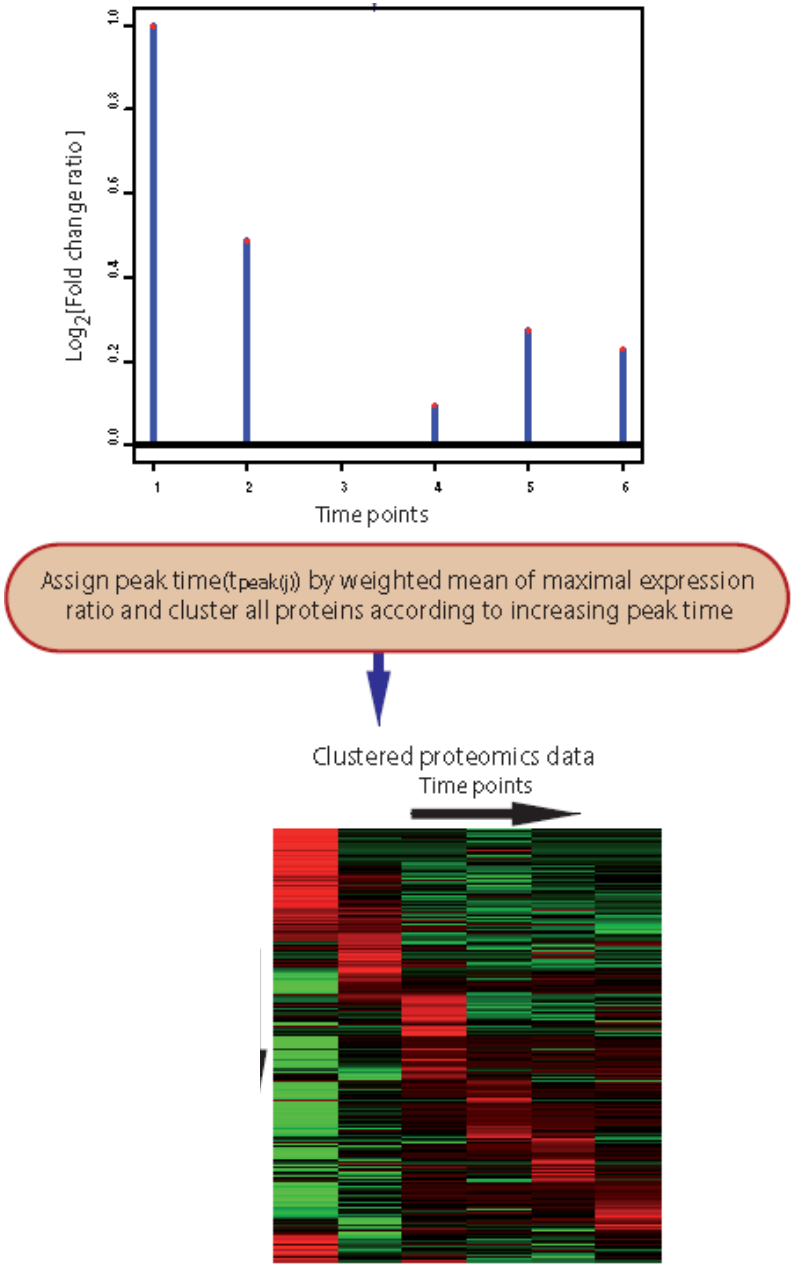
Olsen Science Signaling 3 (2010)

Bioinformatics Workflow (2)



Olsen Science Signaling 3 (2010)

Bioinformatics Workflow (3)

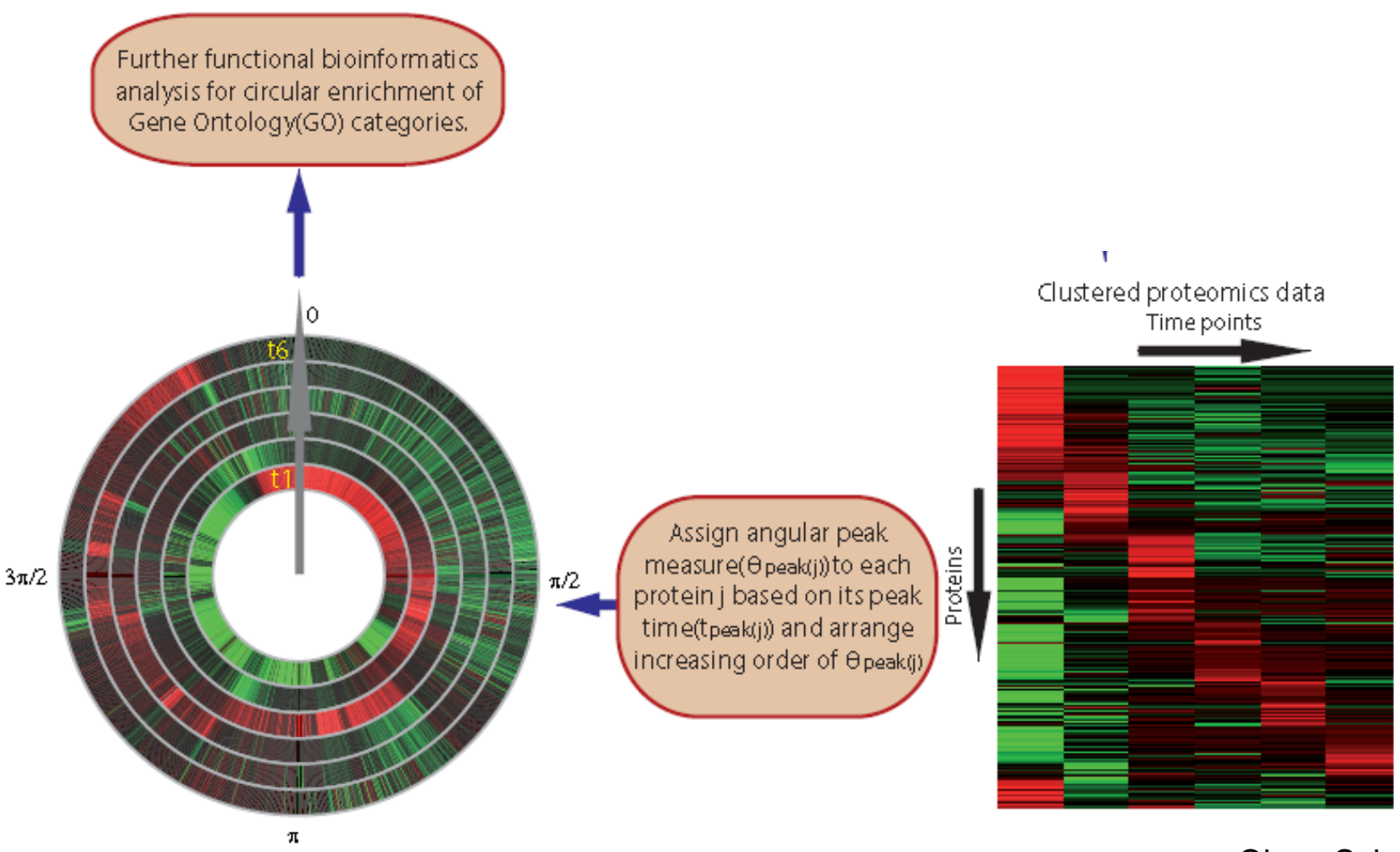


For each protein a **peak time index** was calculated as weighted mean of its maximal expression at time point t_i w.r.t its adjacent time points t_{i-1} and t_{i+1}.

The proteins were then clustered according to increasing peak time indices.

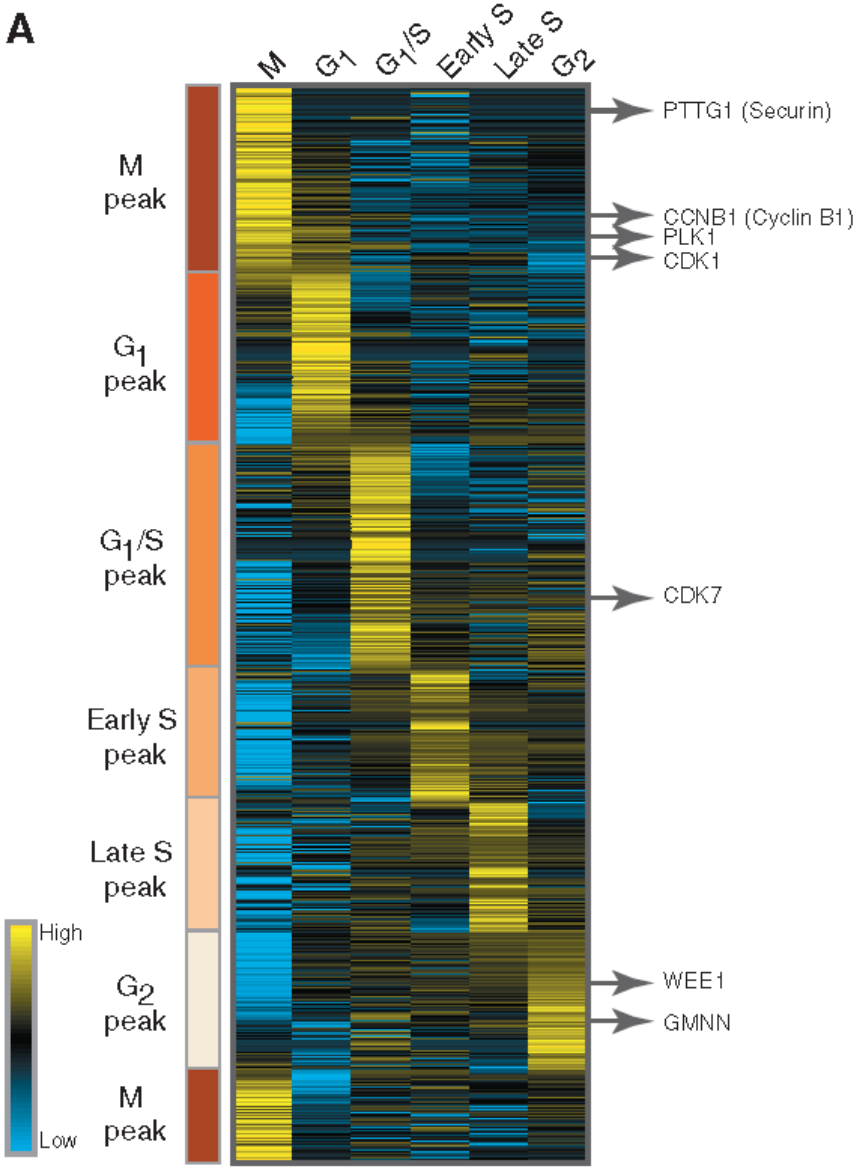
Olsen Science Signaling 3 (2010)

Bioinformatics Workflow (4)



Olsen Science Signaling 3 (2010)

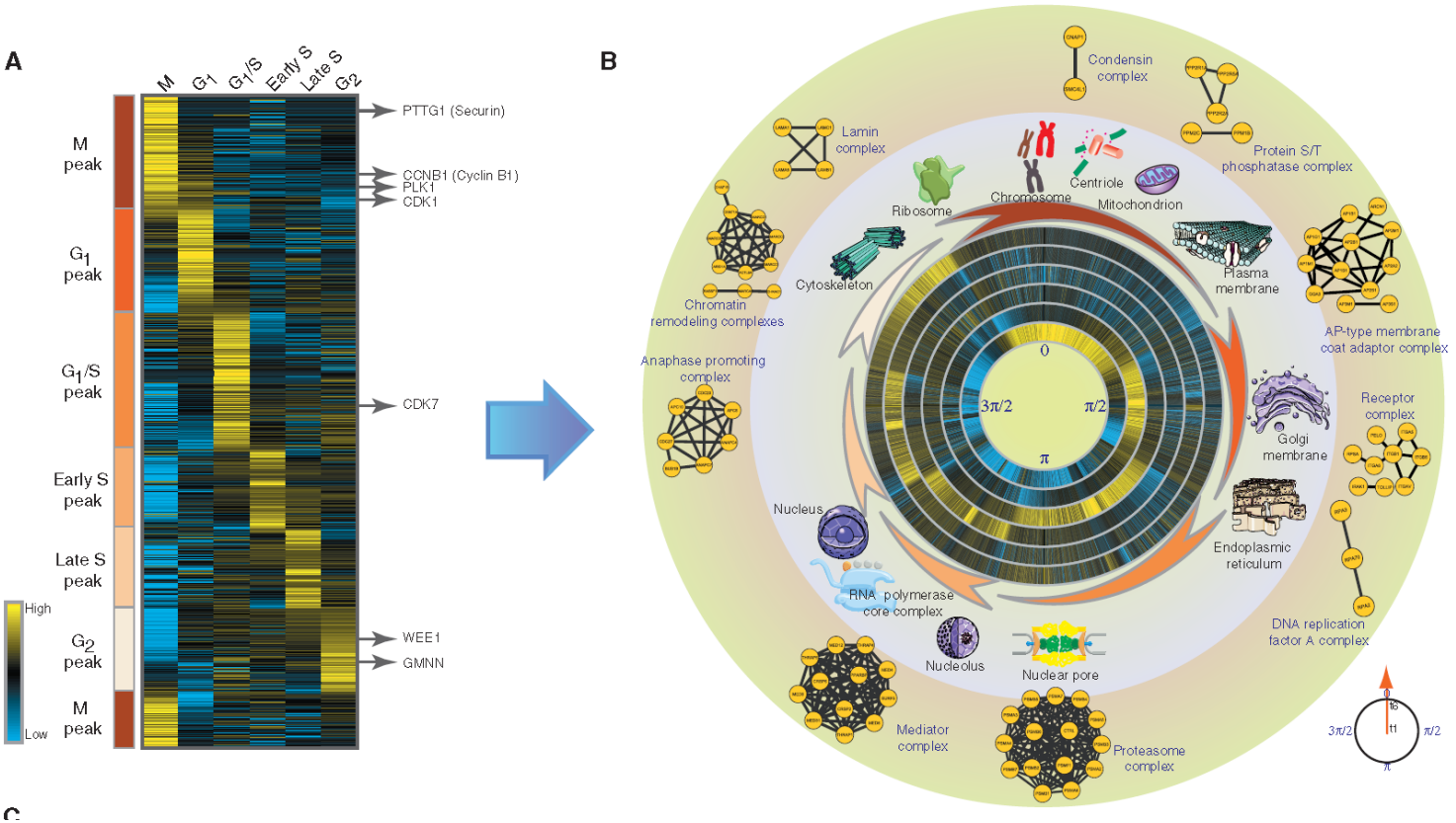
Dynamics of the proteome during the cell cycle



Proteins whose abundance changed at least fourfold during the cell cycle were clustered in all cell cycle stages by calculating a time peak index by weighted mean of the ratio of maximal abundance.

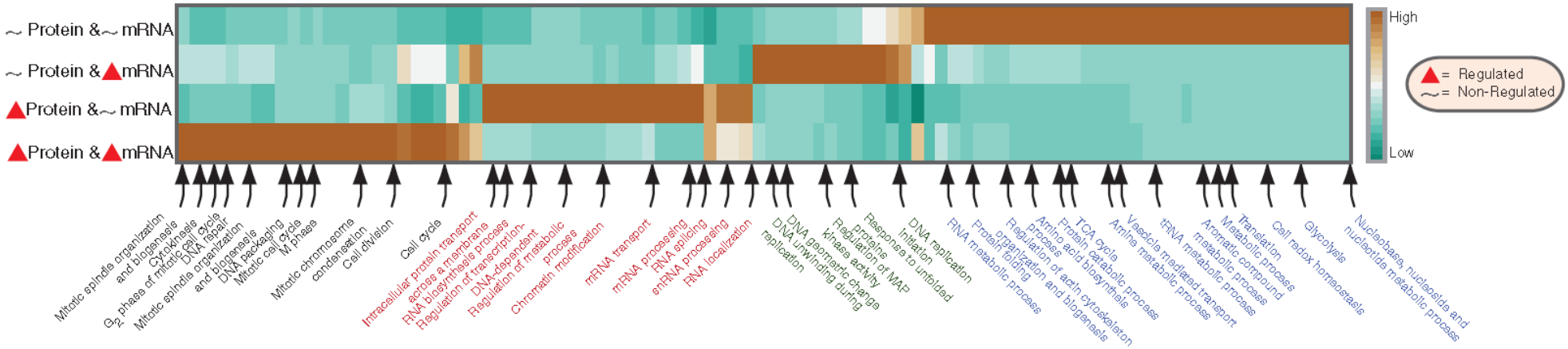
For each cell cycle stage, there are clear patterns of up- and down-regulation.

Determine protein peaks



(B) A circularized representation of the data shown in (A) was used to determine the angle in the cell cycle where the abundance of particular proteins peaks. Coordinately regulated protein complexes and organellar proteins at each cell cycle stage are indicated around the circle.

Comparison of mRNA and protein dynamics



Comparison of mRNA and protein dynamics during the cell cycle.
Measured protein dynamics were correlated to published mRNA data.

Proteins were grouped on the *y* axis in four categories from top to bottom:

- unchanging mRNA and protein
- changing mRNA and unchanging protein
- unchanging mRNA and changing protein
- and changing mRNA and changing protein.

The x axis shows clustered gene ontology (GO) biological process terms enriched in at least one of the above four categories.

High and **low** : statistical over- or underrepresentation.

Absolute phosphorylation site stochiometry

Now we want to derive the phosphorylation state of protein residues during the cell cycle.
We need to substract out the changes of protein abundance.
-> we want to determine (1) and (2) below

(1) Proportion of phosphorylated to unphosphorylated peptide in Light SILAC state: $\frac{N_L^{PHOS}}{N_L^{NonP}} = a$

N_L^{PHOS} is the total copy number of a given phosphopeptide in the light SILAC state, and N_L^{NonP} is the total copy number the corresponding unphosphorylated peptide in the light SILAC state

(2) Proportion of phosphorylated to unphosphorylated peptide in Heavy SILAC state: $\frac{N_H^{PHOS}}{N_H^{NonP}} = b$

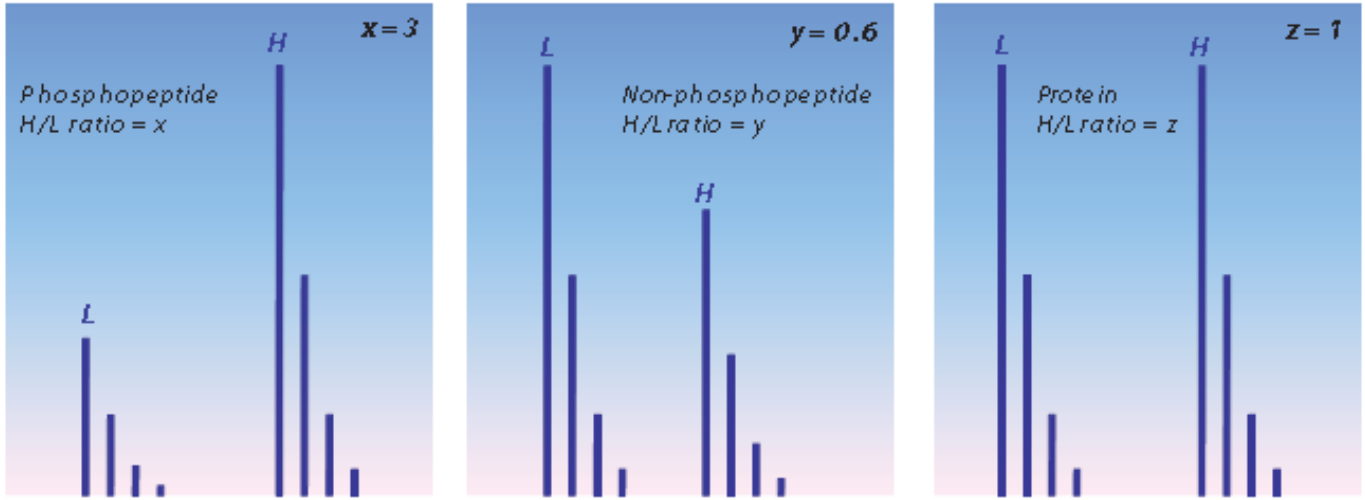
N_H^{PHOS} is the total copy number of a given phosphopeptide in the heavy SILAC state, and
N_H^{NonP} is the total copy number the corresponding unphosphorylated peptide in the heavy SILAC state

(3) We expect that $\frac{N_H^{PHOS} + N_H^{NonP}}{N_H^{PROTEIN}} = \frac{N_L^{PHOS} + N_L^{NonP}}{N_L^{PROTEIN}}$

$N_L^{PROTEIN}$ is the total copy number of the phosphoprotein in the light SILAC state, and
$N_H^{PROTEIN}$ is the total copy number the phosphoprotein in the heavy SILAC state

Olsen Science Signaling 3 (2010)

Available experimental data



To determine phosphorylation sites that show dynamic profiles due to changes in phosphorylation state rather than due to changes in protein abundance, we consider the measured phosphopeptide H/L ratios.

From the experiment we have:

- the SILAC ratio *x* for a particular phosphopeptide
- the SILAC ratio *y* for the respective non-phosphopeptide,
- and protein ratio *z* (the total amount of the protein in both phosphorylated and nonphosphorylated forms).

Absolute phosphorylation site stochiometry

From the MS data we know:

(4) Relative phosphopeptide ratio = $\frac{N_H^{PHOS}}{N_L^{PHOS}} = x$

(5) Relative unphosphorylated peptide ratio = $\frac{N_H^{NonP}}{N_L^{NonP}} = y$

(6) Relative total phosphoprotein ratio = $\frac{N_H^{PROTEIN}}{N_L^{PROTEIN}} = z$

If we know x, y and z then we can solve equations 1 and 2 by substituting in equations 3:

(1) Occupancy rate in Light SILAC state: $\frac{N_L^{PHOS}}{N_L^{NonP}} = a = \frac{z - y}{x - z}$ z

(2) Occupancy rate in Heavy SILAC state: $\frac{N_H^{PHOS}}{N_H^{NonP}} = b = \frac{x \cdot (z - y)}{y \cdot (x - z)}$

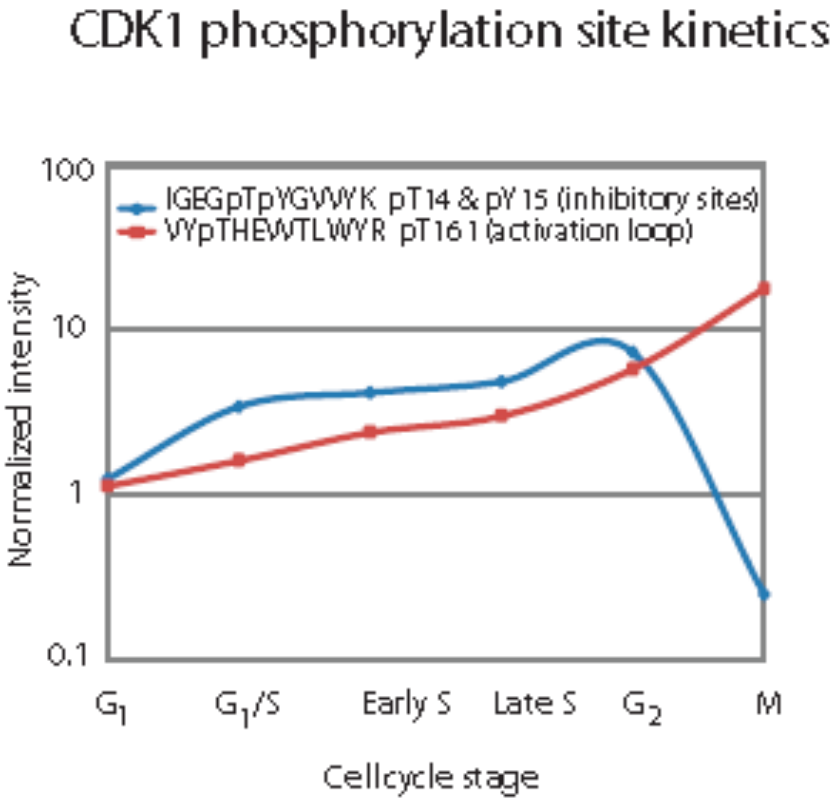
We expect that $N_L^{PHOS} + N_L^{NonP} = N_H^{PHOS} + N_H^{NonP} = 100\% = 1$

and can therefore calculate the phosphorylation site occupancy in the Light and Heavy SILAC state as:

(3) Light SILAC occupancy: a/(1+a) and Heavy SILAC occupancy: b/(1+b)

Olsen Science Signaling 3 (2010)

Example: Dynamic phosphorylation of CDK1

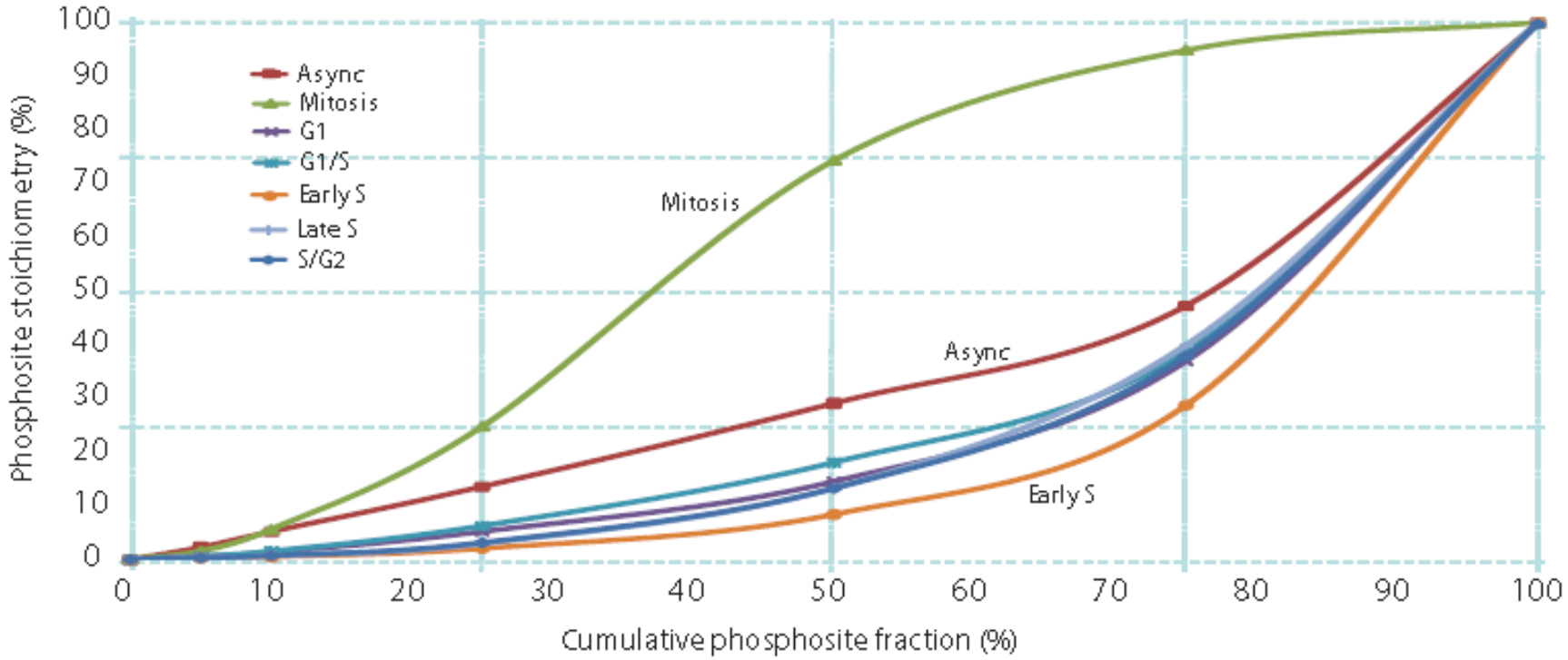


Dynamic profile of two CDK1 phosphopeptides during the cell cycle.

The activating site T161 (**red**) peaks in mitosis, whereas phosphorylation of the inhibitory sites T14 and Y15 (**blue**) is decreased in mitosis

Olsen Science Signaling 3 (2010)

Total phosphosite occupancy in different stages of cell cycle

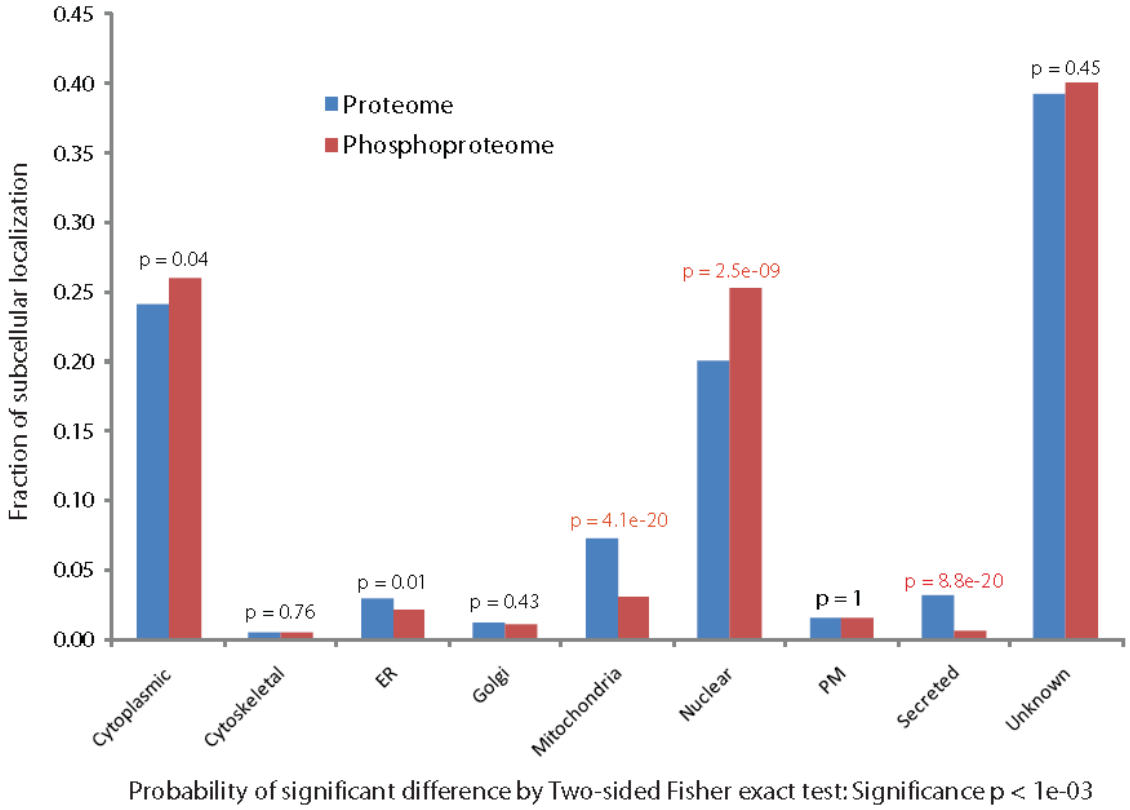


Fifty percent of all mitotic phosphorylation sites have occupancy of 75% or more.

Olsen Science Signaling 3 (2010)

Differential phosphorylation

Gene ontology (GO) analysis of protein and phosphoproteins subcellular localization. All proteins identified by MS were clustered according to their GO annotation for sub-cellular localization (Blue bars). The same clustering was done for all phosphoproteins (Red bars).

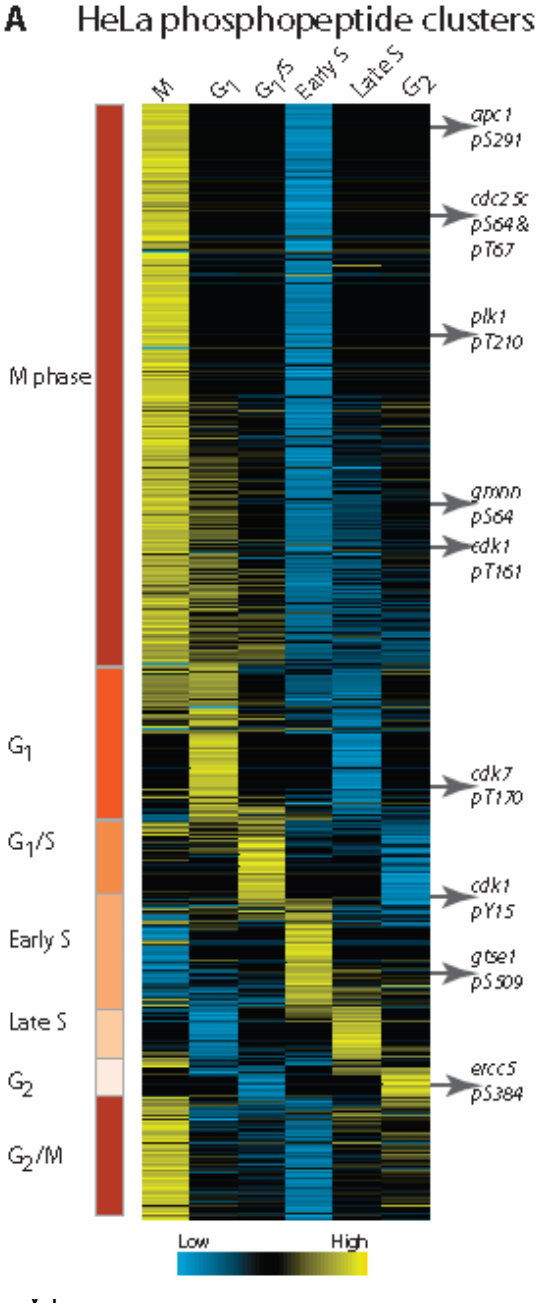


y-axis : percentage of the indicated sub-cellular fractions from the total.

Compared to the proteome distribution, phosphorylated proteins are over-represented in the nucleus and under-represented amongst mitochondrial and secreted proteins.

Olsen Science Signaling 3 (2010)

Dynamics of the Phosphoproteome



Dynamics of the phosphoproteome during the cell cycle.

Clustering of regulated phosphorylation sites in all cell cycle stages.

More than half of all identified regulated phosphorylation sites peak in mitosis.

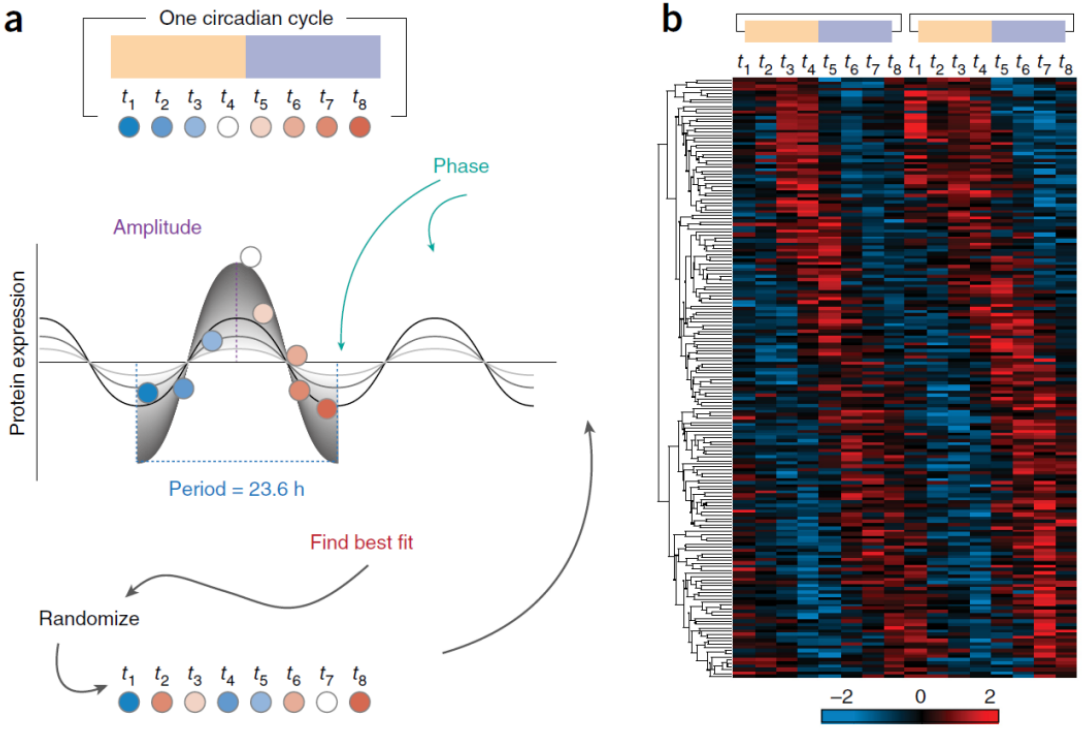
Olsen Science Signaling 3 (2010)

Detect periodic oscillations in time-series analysis

Aim: detect periodic oscillations in protein expression over time.

(a) Amplitude (expression level) and phases (upregulation or downregulation) are determined by optimizing a **cosine function fit** to the data.

A **permutation-based approach** in which the time points are randomly reshuffled multiple times identifies the statistically significantly oscillating proteins, exemplified by global circadian oscillations of the proteome in mouse liver.



(**b**) A total of 180 proteins were found to follow circadian rhythm over two cycles, and characteristic phases of upregulation and downregulation were clearly characterized as illustrated by the red and blue clusters, respectively.

Tyanova et al., Nature Methods 13, 731 (2016)

Data imputation

What is the role of data imputation in MS data?

If no signal is detected, this can have various reasons:

- The peptide is not detected or falsely identified

- The peptide is really not at all present in the sample

- The peptide concentration is below the detection threshold …

The reason for missing data is generally not known.

Simply setting all missing data to zero would generate **false positive** signals
= proteins appear to be significantly deregulated, but are in fact not.

Imputation methods: KNNimpute

Lets assume that gene $\mathbf{g}_1$ lacks data point *i*.

The KNNimpute method (Troyanskaya *et al.*, 2001) finds *k* (*k* < *m*) other genes with expressions most similar to that of $\mathbf{g}_1$ and that do have a measured value in position *i*.

The missing value of $\mathbf{g}_1$ is estimated by the weighted average of the values in position *i* of these *k* closest genes.

$$\mathbf{g}^* = \frac{\omega_1 \mathbf{g}_{s_1} + \omega_2 \mathbf{g}_{s_2} + \cdots + \omega_k \mathbf{g}_{s_k}}{\omega_1 + \cdots + \omega_k},$$

Here, the contribution of each gene is weighted by the similarity of its expression to that of $\mathbf{g}_1$.

$$\omega_i = 1 / \| \mathbf{w} - \mathbf{a}_i \|_2,$$

Kim et al., Bioinformatics 21, 187 (2005)

Imputation methods: SVDimpute

SVDimpute method (Troyanskaya *et al.*, 2001):

- Given: matrix G where some data is missing.
- Generate initial matrix *G'* from G by substituting all missing values of the *G* by zero or row averages.
- Compute SVD of G'.
- Determine the *t* most significant eigengenes of *G'* (with largest eigenvalues).
- Regress every gene with missing values against the *t* most significant eigengenes (by ignoring position *i*)

Using the coefficients of the regression, the missing value in G is estimated as a linear combination of the values in the respective position *i* of the *t* eigengenes.

This procedure is repeated until the total change of the matrix G becomes insignificant.

Kim et al., Bioinformatics 21, 187 (2005)

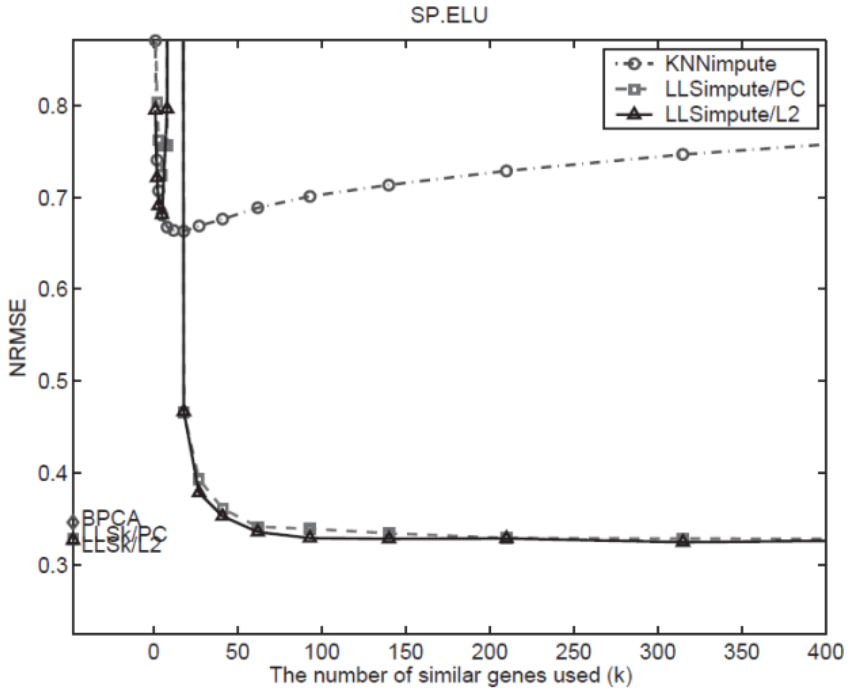
Imputation methods: Local Least squares

(1) select *k* genes that have similar properties (e.g. expression profiles) to the gene with missing information based on the *L2*-norm or Pearson correlation coefficients of the expression profiles.

(2) regression and estimation, regardless of how the *k* genes are selected.

Spellman data set: yeast cell cycle
5% of data were missing

-> LLSimpute outperforms KNNimpute



Kim et al., Bioinformatics 21, 187 (2005)

Imputation methods: Local Least squares

Based on the k-neighboring gene vectors, form the matrix $A \in \mathrm{R}^{k \times (n-1)}$ and the two vectors $\mathbf{b} \in \mathrm{R}^{k \times 1}$ and $\mathbf{w} \in \mathrm{R}^{(n-1) \times 1}$.

The k rows of the matrix A consist of the k-nearest neighbor genes $\mathbf{g}^{\mathrm{T}}_i \in \mathrm{R}^{1 \times n}$, $1 \le i \le k$, with position i deleted.

The elements of the vector **b** consists of position i of the k vectors $\mathbf{g}^{\mathrm{T}}_i$.
The elements of the vector **w** are the $n - 1$ elements of the
gene vector $\mathbf{g}_1$ whose missing position is deleted.

After the matrix A, and the vectors **b** and **w** are formed, the least squares problem is formulated as

$$\min_{\mathbf{x}} \| A^{\mathrm{T}}\mathbf{x} - \mathbf{w} \|_2$$

Then, the missing value α is estimated as a linear combination of the respective values of the neighboring genes

$$\alpha = \mathbf{b}^{\mathrm{T}}\mathbf{x} = \mathbf{b}^{\mathrm{T}}(A^{\mathrm{T}})^{\dagger}\mathbf{w}$$

Kim et al., Bioinformatics 21, 187 (2005)

Models for missing values

Missing Completely At Random (MCAR): in a proteomics data set, this corresponds to the combination of a propagation of multiple minor errors or stochastic fluctuations. e.g. by a misidentified peptide

Missing At Random (MAR): this is a more general class than MCAR, where conditional dependencies are accounted for. In a proteomics data set, it is classically assumed that all MAR values are also MCAR.

Missing Not At Random (MNAR) assumes a targeted effect. E.g. in MS-based analysis, chemical species whose abundances are close enough to the limit of detection of the instrument record a higher rate of missing values.

Imputation methods for MCAR and MAR are general.
For MNAR, they are methods-specific.

Lazar et al., J Proteome Res 15, 1116 (2016)

Simulation benchmark

Use real data (Super-SILAC and label-free quantification) on human primary tumor-derived xenograph proteomes for the two major histological nonsmall cell lung cancer subtypes, adenocarcinoma and squamous cell carcinoma, using.

MNAR values: one randomly generates a threshold matrix T from a Gaussian distribution with parameters ($\mu_t = q$, $\sigma_t = 0.01$), where q is the α-th quantile of the abundance distribution in the complete quantitative data set.
Then, each cell (i,j) of the complete quantitative data set is compared with $T_{i,j}$.
If $(i,j) \geq T_{i,j}$, the abundance is not censored.
If $(i,j) < T_{i,j}$, a Bernoulli draw with probability of success $\beta\alpha \cdot 100$ determines if the abundance value is censored (success) or not (failure).

MCAR values are incorporated by replacing with a missing value the abundance value of n m*((100 - β) α /100) randomly chosen cells in the table of the quantitative data set.

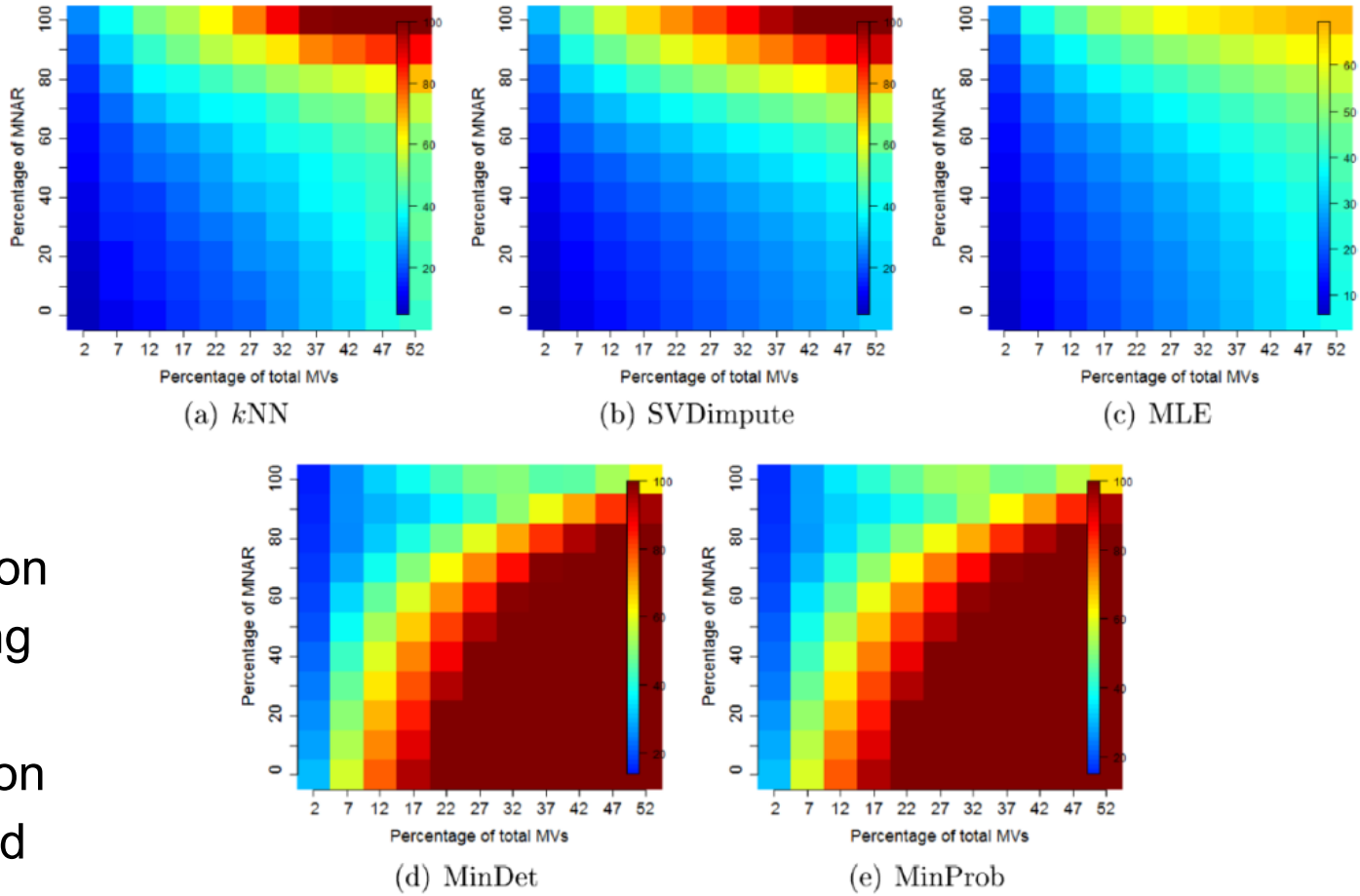
Lazar et al., J Proteome Res 15, 1116 (2016)

Imputation methods: benchmark

MLE: maximum likelihood estimator

MinDet: simply replace missing values by the minimum value that is observed in the data set.

MinProb: stochastic version of MinDet. Replace missing values with random draws from a Gaussian distribution centered on the value used with MinDet and with a variance tuned to the median of the peptide-wise estimated variances



RSR = RMSE / std.dev.

Blue: low RSR

Red: high RSR

Lazar et al., J Proteome Res 15, 1116 (2016)

Conclusion on data imputation

Algorithms SVDimpute, kNN, and MLE perform better under a small MNAR ratio.

Algorithms MinDet and MinProb better under a larger MNAR ratio.

Algorithms of the first group generally seem to give better predictions.

Kim et al., Bioinformatics 21, 187 (2005)