V3 – 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
- **Data imputation** for MS data
  - Identify TRAP clients

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

V3 Processing of Biological Data SS 2020 1

In lecture 3, we will deal with data on protein expression levels. Nowadays, these data are typically determined by mass spectrometry.

First, we will review some basics about the mass spectrometry methods.

Then, we will turn at bioinformatics tasks in processing MS data.

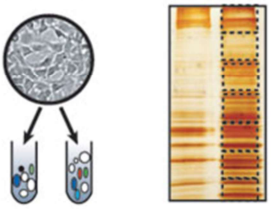
Phosphorylation is a very important post-translational modification. MS is the ideal method to dectect site-specific phosphorylation.

Finally, we will turn to a collaboration project between our group and that of Prof. Richard Zimmermann from the medical department in Homburg.

Proteomics workflow: (1) protein isolation

(1) Sample fractionation

SDS–PAGE



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.**

The mass of the intact protein by itself is insufficient for identification.

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

V3 Processing of Biological Data SS 2020 2

The application of mass spectrometry to study proteins became popular in the 1980s after the development of the MALDI and ESI techniques.

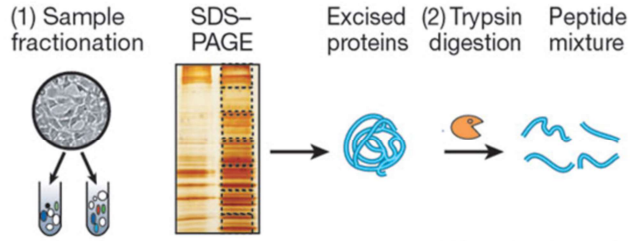
ESI stands for electrospray ionization, MALDI for matrix-assisted laser desorption/ionization.

They are the two primary methods used for the ionization of protein in mass spectrometry.

John B. Fenn and Koichi Tanaka made crucial contributions to the development of ESI and MALDI, respectively, and received the Noble prize for this.

The first stage of a proteomics experiment does not involve a mass spectrometer yet. First one needs to isolate the proteins of interest from the biological sample.

Proteomics workflow: (2) trypsin digestion



Therefore, in stage 2, proteins are **degraded enzymatically** to peptides, usually by trypsin.

This yields peptides with C-terminally protonated amino acids (K/R) which is beneficial 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)

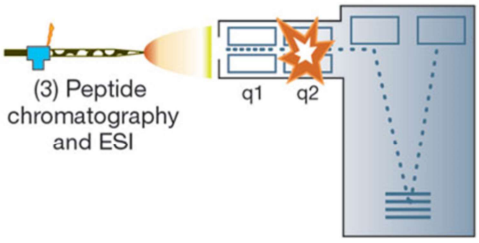
Henzel et al. J Am Soc Mass Spectrom
14, 931–942 (2003)

The second stage consists of digesting the purified proteins with a suitable enzyme.

As listed in table 1, the enzyme trypsin cleaves peptide chains at the positively charged amino acids lysine or arginine.

This typically generates short peptide fragments of around 8 amino acids in length.

Proteomics workflow: (3) peptide chromatography



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)

The third stage typically consists of a chromatography step and the generation of the ionized fragments.

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)

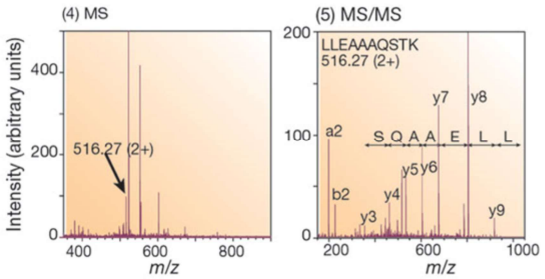
These are notes on the principles of ESI and MALDI.

Proteomics workflow: (4) MS

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 then 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)

V3 Processing of Biological Data SS 2020 6

The mass spectrometer detects mass over charge ratios (m/z).

Panel (4) shows 2 high peaks surrounded by many small peaks.

In this example, a smaller peak marked by an arrow and labeled 516.27 (2+) is selected.

516.27 stands for its mass in Dalton units. (Remember, a Dalton is defined as 1/12 of the mass of an unbound neutral atom of carbon-12).

2+ is the charge of this peptide fragment in electron charges.

The molecules collected under this peak are sent into the mass spectrometer again („tandem mass spectrometer").

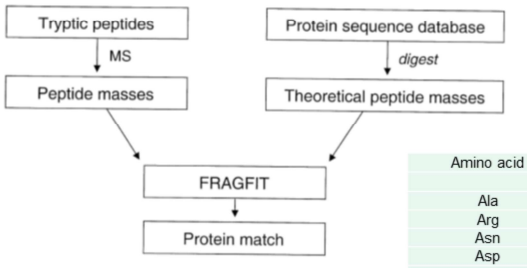
Panel (5) shows the fragments detected for the peptide LLEAAAQSTK.

One can detect many fragments at different m/z values.

Assuming that all carry the same net charge, one can associate the distances between the peaks to the mass differences between peptide fragments of different length.

As shown here, one can identify fragments matching the peptide sequence.

Peptide mass fingerprinting



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

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

V3 Processing of Biological Data SS 2020 7

The mass of a peptide fragment can simply be computed by summing up the masses of its building blocks, the amino acids.

By matching the identified peptide fragments to protein sequences in a database, one can identify the protein that was originally purified from the sample.

How many peptides are detected?

There are several reasons why an analysis does not find all amino acids.

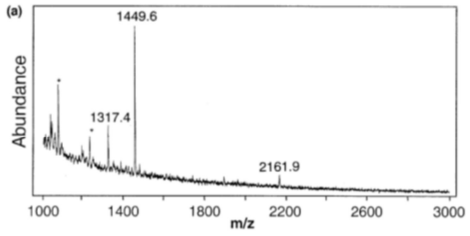
- protein does not digest well
- peptides too hydrophilic or small-they pass through the reverse phase column with salt and are not analyzed
- peptides too large/hydrophobic-they stick in gel, adsorb to tubes, do not elute from column, or are too large for the mass spectrometer to analyze because of poor fragmentation
- peptides fragment in ways which cannot be analyzed. Many spectra in an analysis cannot be interpreted. Some spectra only give limited data; proline, histidine, internal lysine and arginine are some reasons peptides do not give complete fragmentation data.

Seeing enough peptides to show 70% of the sequence of a protein (70% coverage) is a very successful protein analysis.

https://med.virginia.edu/biomolecular-analysis-facility/services/mass-spectrometry/protein-analysis-by-mass-spectrometry/

These are notes from a mass spectrometry service facility at the University of Virginia.

Peptide mass fingerprinting



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

3 characteristic peaks are labeled.

(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

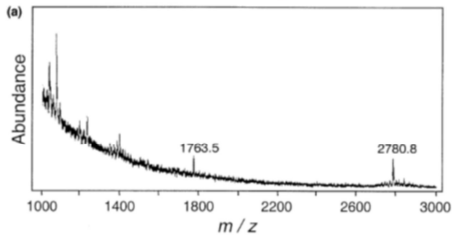
Mass [Da]	Starting position	Peptide fragment
2162.444	84:	DGRTPGSRNLCNIPCSALLSS
1449.706	105:	DITASVNCAKIVS
1317.552	137:	DVQAWIRGCRL

(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)

Identifying matching peptides derived from the protein lysozyme in the protein sequence database.

Peptide mass fingerprinting



(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.

```
(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
```

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)

Identifying matching peptides in the protein sequence database.

In higher organisms, the sequence of cytochrome c is usually 104 amino acids long.

Two peptides of 15 AA and 24 AA in length were sufficient to identify protein and species.

Apparently, this technique did not use trypsin digestion but CNbr which produces fragments of average length 38 AA.

Then, one needs of course fewer peptides.

Application of MS: 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** by:
- influencing the tertiary and quaternary **structure** of a protein,
- controlling **subcellular distribution**, and
- regulating its **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)

No comments.

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 human proteins get phosphorylated.

This study has been cited more than 1200 times.

The authors monitored phosphorylation of proteins during the cell cylce of HeLa cells.

They found that about 70% of all human proteins get phosphorylated, on average in 3-4 different sites.

Note that phosphorylation often determines the activity of the protein.

The dynamics of protein levels and phosphorylation levels was determined with the SILAC method.

Review: protein quantification by SILAC

ARTICLE

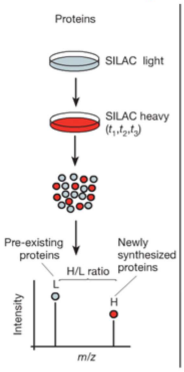
doi:10.1038/nature10098

Global quantification of mammalian gene expression control

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

SILAC: „stable isotope labelling by amino acids in cell 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)

In the SILAC method, cells are first grown in a normal medium, which is then supplemented by heavy isotope-versions of essential amino acids.

Essential amino acids are those that the cells cannot make themselves and need to uptake from the medium.

After exchanging the medium, the cells continue to synthesize proteins, now using the heavier versions of the amino acid building blocks.

Thus, the sample will then contain „light“ copies of each protein (labeled L) that pre-existed when the medium was exchanged and new „heavy“ copies (labeled H).

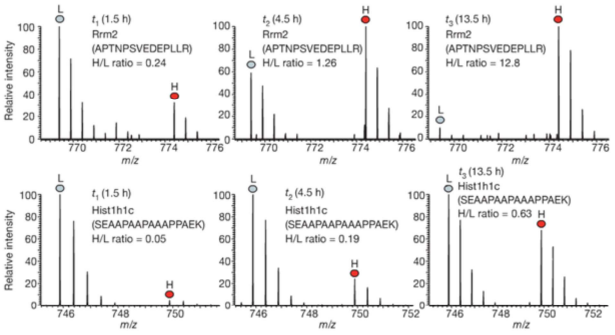
Rates of protein translation

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)

Shown is the time-evolution of the concentration of „light“ and „heavy“ peptides.

The upper row shows a high-turnover peptide from the Rrm2 protein.

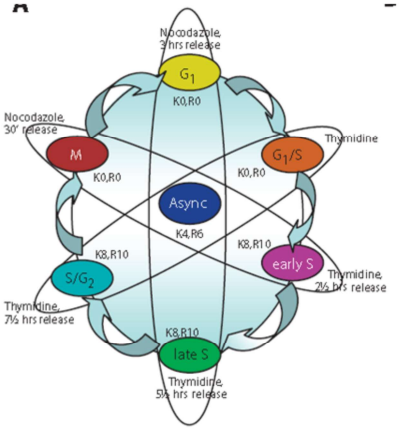
Between time points 1.5 hours (left) and 4.5 hours (middle), the number of L copies has decreased from over 100 to about 70 and the number of H copies from 40 to over 100.

The increase of H reflects the synthesis of new proteins. The decay in L reflects the exponential decay of the pre-existing copies with a characteristic (fast) half-time.

After 13.5 hours, the number of H copies has remained the same as after 4.5 hours, showing that synthesis and decay are now balanced.

The bottom row shows the same process for a low-turnover peptide that grows slower (H form) and also decays much slower (L form).

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 6 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.

V3 Processing of Biological Data SS 2020 Olsen Science Signaling 3 (2010) 15

The authors applied 2 molecules that cause cell cycle arrest at different stages, thymidine and nocodazole.

Thymidine blocks entry into S phase. Nocodazole arrests cell during mitosis.

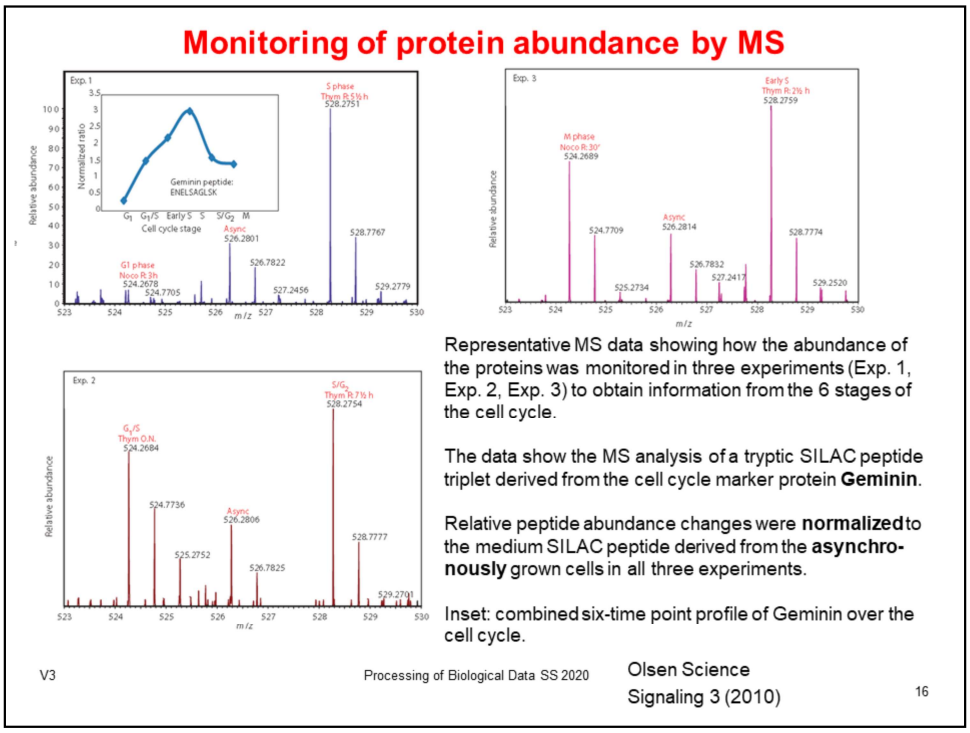
In this way, all cells can by synchronized at one stage of the cell cycle.

By washing steps, one can wash out the molecules and restart cell cycle.

In the figure, this is marked as „release“ = cells are released from arrest.

To save costs, the authors always mixed 3 cell populations that are marked here by ellipsoids and that were grown with different SILAC-labels.

All experiments contain the „async“ sample – this can then be used to normalize the protein levels from different experiments.



The 3 panels show 3 experiments for a short peptide ENELSAGLSK derived from the cell cycle marker protein Geminin.

As explained before, each panel contains data from 2 opposite cell cycle stages and from the „async“ mixture.

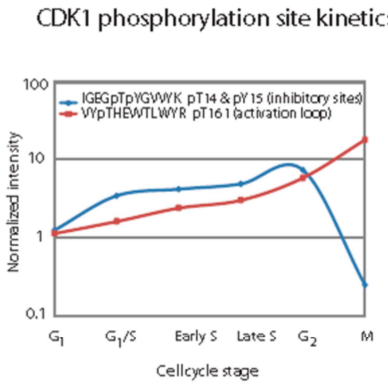
„Async“ is always placed in the middle of the x-axis – meaning that it was always labeled with the medium SILAC label.

Peaks on the right have heavier masses and were labeled by heavy SILAC label.

Each spectrum contains a set of peaks („fingerprint“) that are characteristic for this peptide.

By combining the data from different panels, and normalizing the data, one obtains the expression profile of this peptide during the cell cycle shown in the inset of the top left panel.

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)

The figure shows the levels of 2 phosphopeptides belong to the CDK1 kinase during the cell cycle.

Phosphorylation of Thr161 increases during the cell cycle, that of Thr14 and Tyr15 sharply decreases in mitosis.

Data imputation for MS

What is the role of data imputation in MS data?

It is not unusual for LC–MS proteomics datasets to have **as much as 50% missing peptide values** (J Proteome Res. 14, 1993 (2015)).

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.

Dealing with missing values is a major task when processing data from mass spectrometry.

On slide 8, we have listed possible reasons why certain peptides are not detected at all.

But this does not explain why they can be detected in one sample, but not in another one.

We will not go deeper into this here. It is sufficient for you to realize the enormous importance of this point.

Imputation methods: KNNimpute

Lets assume that gene $\mathbf{g_1}$ lacks data point *i* (for condition *i* or for patient *i*) and the total number of genes is *m*.

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 $\mathbf{g_1}$.

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

V3 Processing of Biological Data SS 2020 19

KNN stands for k nearest neighbors.

The idea follows the often used principle „guilt by association“.

If k other genes show a very similar expression profile to gene1 under all (or many) other conditions, then it makes sense to impute the missing expression level of gene1 based on the values of the other genes in condition i.

The formula shows that a weighted schema is used, where the weights represent the similarity of expression to gene1.

Obviously, we can apply this algorithm unchanged to protein levels instead of mRNA levels.

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 *G* by 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)

V3 Processing of Biological Data SS 2020 20

SVD can only be performed on complete matrices.

Therefore, a second matrix G' is constructed where all missing values are replaced by row averages.

SVD yields all eigenvectors. Those with largest eigenvalues are termed eigengenes.

Then, we compute for each gene (here: protein) the coefficients of a linear combination of the leading eigengenes.

For this, we can only use the known data points.

The missing data point is then computed with the same linear combination.

With these imputed data points, we can compute new row averages, and redo the SVD of G'.

Imputation methods: Local Least squares

(1) select *k* genes that have similar properties (e.g. expression profiles) to the gene where position *i* is missing.

Similarity can be based on the *L*2-norm (same as Euclidian norm)

$$d(\mathbf{p},\mathbf{q}) = \sqrt{\sum_{i=1}^{n}(q_i - p_i)^2}$$

or Pearson correlation coefficients of the expression profiles.

(2) regression and estimation

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

The local least squares method for data imputation combines elements from the kNNimpute and SVDimpute methods.

Again, one uses information of genes with similar expression (identified either by L2 norm or Pearson correlation).

Imputation methods: Local Least squares

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

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

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

After the matrix A, and the vectors $\mathbf{b}$ and $\mathbf{w}$ are formed, the least squares problem is formulated as

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

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

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

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

V3 Processing of Biological Data SS 2020 22

Matrix A contains the expression profiles of the k nearest genes.

Vector b contains their expression values at the missing position i.

Vector w contains the expression values of gene1 except the missing position i.

One finds a vector x (stands for a linear combination of the other genes) so that A^{T}x is as close as possible to w.

Explanation: x projects the expression values of the other genes onto the expression of gene1.

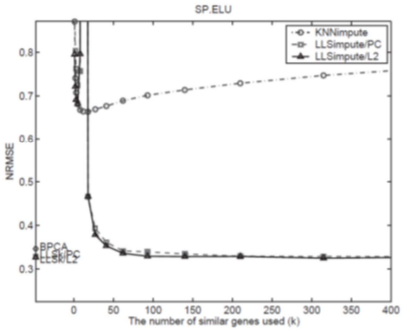
Then one can also use this vector x to project the data points for the other genes in i onto gene1. This is done in the last formula here.

Imputation methods: Local Least squares

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

-> LLSimpute outperforms KNNimpute

Lower Root Mean Square Error (RMSE)



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

The experiment of Spellman et al. is described in a classic paper (https://www.ncbi.nlm.nih.gov/pubmed/9843569). The authors used microarrays to identify periodically cycling genes along the cell cycle of yeast.

Shown on the x-axis is the number of neighboring genes used.

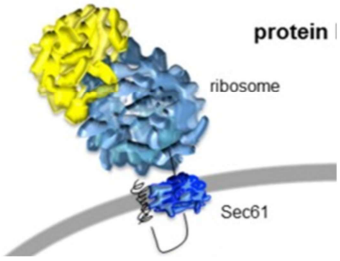
For kNN, there is an optimal number of maybe 10 genes, then the deviation from the correct data points increases again.

LLSimpute shows about twice as good results as kNN (RMSE is less than half) and converges for arbitrarily many genes used.

Case study: identify clients of TRAP complex

In mammalian cells, one-third of all polypeptides are transported into or across the ER membrane via the Sec61 channel.

The Sec61 complex facilitates translocation of all polypeptides with signal peptides (SP) or transmembrane helices.



The Sec61-auxiliary translocon-associated protein (**TRAP**) complex supports translocation of only a **subset of precursors**.

To characterize determinants of TRAP substrate specificity, we here systematically identify TRAP-dependent precursors by analyzing cellular protein abundance changes upon siRNA-induced TRAP depletion by proteome MS.

Lang et al. Front Physiol. (2017) 8: 887

This is an example from our own work on proteomic data.

The group of Prof. Richard Zimmermann from Homburg has studied the Sec61 complex since more than 30 years.

The Sec61 has an important role for protein synthesis.

There exist two sorts of ribosomes, cytosolic ribosomes and ribosomes that bind to the membrane of the endoplasmic reticulum.

Cytosolic ribosomes synthesize cytosolic proteins. We will not consider this here.

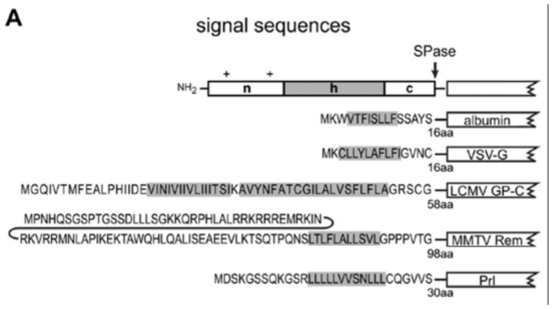
ER bound ribosomes synthesize membrane proteins and proteins that will be excreted by the cell via exocytosis.

Once the nascent peptide chain leaves the ribosome tunnel, it enters the pore of the Sec61 complex and is either released into the membrane or translocated inside the ER.

However, some proteins cannot translocate by themselves, they require the activity of accessory membrane proteins.

Signal peptides

Examples of differently sized signal sequences.
Signal sequences can be as small as 16 amino acid residues but some are more then 50 amino acid (aa) residues in length.
A characteristic feature of a signal sequence is its hydrophobic (h) region. Examples of minimal (albumin and VSV-G protein) and extended signal sequences (LCMV GP-C, MMTV Rem and prolactin (Prl) are shown.



Günter Blobel, Noble prize 1999 "for the discovery that proteins have intrinsic signals that govern their transport and localization in the cell." He donated his prize money to support rebuilding the Frauen-kirche in Dresden.

Kapp, Katja; Schrempf, Sabrina; Lemberg, Marius K.; Dobberstein, Bernhard. *Post-Targeting Functions of Signal Peptides*. Landes Bioscience.

Shown on the top right are several signal sequences.

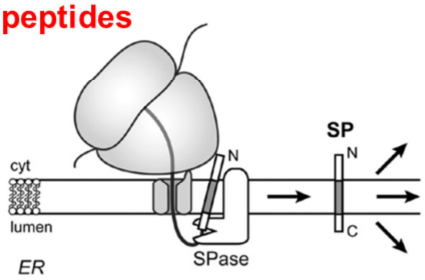
They typically contain an N-terminal „n“ region with several positvely charged amino acids, a hydrophobic „h“ region, and a polar C-terminal „c“ region.

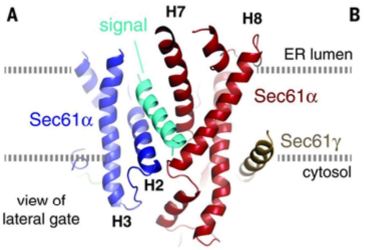
One of the discoverers of signal sequences, Günter Blobel, received the Nobel prize for this.

Signal peptides

After insertion into the Sec61 complex in the ER membrane, signal sequences are usually cleaved off by **signal peptidase** (SPase) on the lumenal side of the ER membrane.

The resulting signal peptides (SPs) initially accumulate in the ER membrane. Subsequently they can become degraded or can have functions as membrane-integrated peptides or as peptides released from the membrane either into the cytosol or the ER lumen.





Kapp, Katja; Schrempf, Sabrina; Lemberg, Marius K.; Dobberstein, Bernhard. *Post-Targeting Functions of Signal Peptides*. Landes Bioscience.

Voorhees, R. and Hegde, R. S. (2016) Structure of the Sec61 channel opened by a signal sequence. Science 351: 88-91.

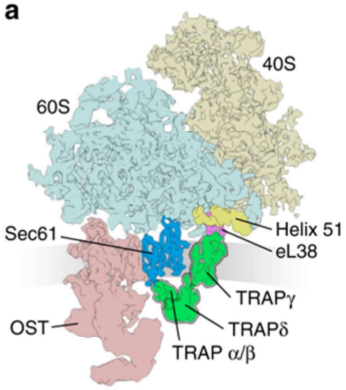
The signal peptide inserts into the Sec61 complex and then somehow turns around.

The Sec61 complex opens a lateral gate (bottom figure, X-ray structure), the SP

is cleaved by the ER-enzyme signal peptidase, and partitions into the membrane.

Ribosome : Sec61 : TRAP : OST supracomplex

a



Cartoon of clipped 80S ribosome together with Sec61-complex (**blue**), TRAP-complex (**green**), and Oligo Saccharyl Transferase.

Structure determined by cryo-EM

AIM:
We tried to determine which proteins can translocate through Sec61 alone and which ones require the help of Trap.

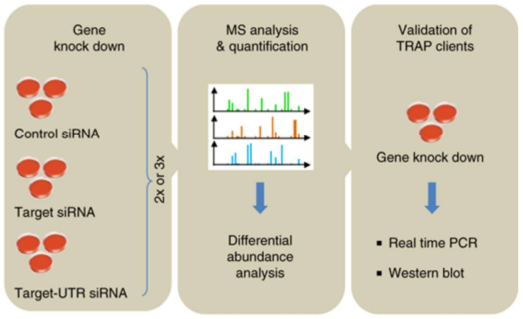
Duy et al., Nature Commun 9, 3765 (2018)

Stefan Pfeffer and Friedrich Förster (MPI Martinsried) were able to detect the structure of ribosomes bound to the Sec61complex by CryoEM.

They could also annotate electron density to the enzyme oligo saccharyl transferase that adds sugar units to the translocated proteins

and to subunits of the TRAP complex.

Experimental strategy



siRNA-mediated gene silencing using two different siRNAs for each target and one non-targeting (control) siRNA, respectively.

6/9 replicates for each siRNA in 2/3 independent experiments.

Label-free quantitative proteomic analysis and differential protein abundance analysis identify negatively affected proteins (i.e., clients) and positively affected proteins (i.e. compensatory mechanisms).

Duy et al., Nature Commun 9, 3765 (2018)

V3 Processing of Biological Data SS 2020 28

This is the experimental strategy to identify which proteins require Sec61 and accessory proteins for translocation.

The main strategy is to knock-down synthesis of new Sec61 or new accessory proteins by siRNA.

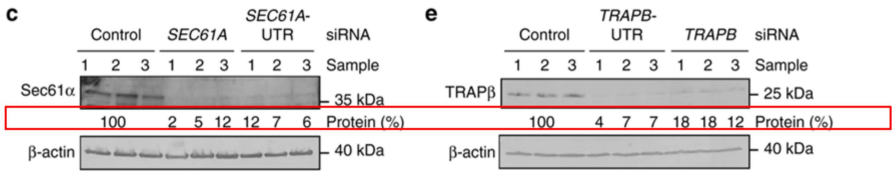
Then, MS is used to identify proteins in the cell lysate (middle lane).

Our task was to identify differentially abundant proteins between samples of two types (i.e. with and without siRNA silencing).

Validation of knock-down

Knock-down of SEC61-alpha

TRAP-beta



Knock-down efficiencies were evaluated by western blot.

Results are presented as % of residual protein levels (normalized to ß-actin) relative to control, which was set to 100%.

Q: why do the levels of SEC61 and TRAP do not go to zero after siRNA silencing (for 72 – 96 hours)?

This slide shows that Sec61alpha levels (left) and TRAPbeta subunit levels (right) were silenced to a few percent. This confirms that silencing worked well.

Although silencing was carried out over 4 days, some residual Sec61alpha or TRAPbeta protein was still left.

This is actually quite good and avoids that the cells may die.

The lower lines show the protein levels of beta-actin. This is a cytoskeletal protein, which should always be there at similar levels.

Experimental silencing strategy

Each MS experiment provided proteome-wide abundance data as LFQ intensities (Cox et al. Mol Cell Proteomics. (2014)13: 2513–2526 – how to combine peptide intensities into aggregated protein abundances?)

for 3 sample groups :

- one control (non-targeting siRNA treated) and
- two stimuli (down-regulation by two different targeting siRNAs directed against the same gene)

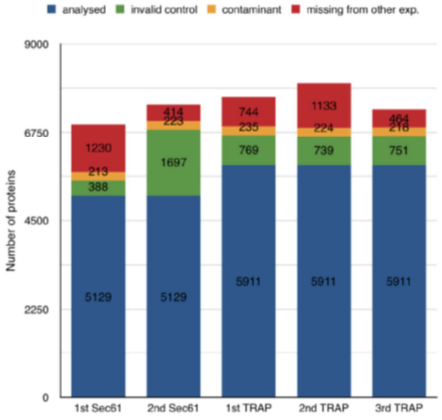
each having 3 data points.

For each sample, 3 replicate experiments were performed.

The control sample is a sample treated with an siRNA that does not target Sec61 and presumably no other gene.

Then, there are 2 samples from silencing experiments where two different siRNA molecules were used.

Number of proteins in MS experiments



Number of proteins detected in the 2 Sec61 depletion experiments (two left most columns) and in the 3 TRAP depletion experiments (three rightmost columns).
Blue bars : proteins analyzed here.
Green : proteins that do not have sufficient control data points, i.e. more than 2/3 of the control samples have missing data points.
Yellow : "contaminants" from MaxQuant analysis.
Red : proteins that cannot be found (or contain "invalid control") in other corresponding experiments.

The number of proteins detected in Sec61 and TRAP silencing experiments was 7212 ± 356 and 7670 ± 332, respectively (mean values with standard deviation, n=2 and n=3, respectively). The observed difference of about 460 was just a bit outside of the standard deviation and is, hence, not statistically significant.

V3 Processing of Biological Data SS 2020 31

In the MS experiments, between 6800 – 8000 proteins were detected. These are typical numbers for such experiments.

We omitted 3 classes of proteins from this dataset:

„red" cases are proteins that are not found in the other experiments

„yellow" cases are proteins classified as contaminants by the MaxQuant software

„green" cases are proteins that were not detected in any of the 3 control replicates.

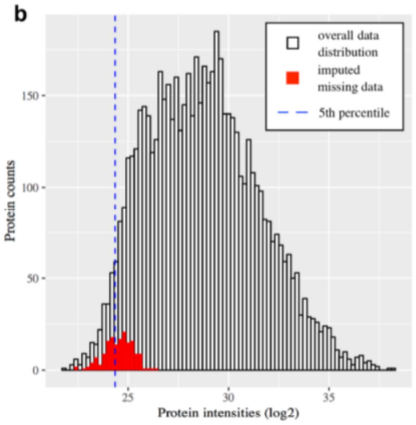
This means we considered 5129 proteins for the Sec61 silencing experiment and 5911 proteins for the Trap silencing experiment.

Imputation strategy – no valid data points

Missing data points were generated by imputation. We distinguished 2 cases.

For **completely missing proteins** lacking any valid data points after siRNA knock-down, imputed data points were randomly generated in the bottom tail of the whole proteomics distribution.

This is based on the assumption that they come from proteins which have limited number of copies that cannot be detected by the mass spectrometer.



As explained on the previous slide, we omitted cases which did not have any non-zero abundance measurement for the control samples.

However, we kept cases that have zero abundance in all silencing experiments.

In that case, we applied the standard strategy used by the Perseus software (https://www.nature.com/articles/nmeth.3901) from the MPI in Martinsried.

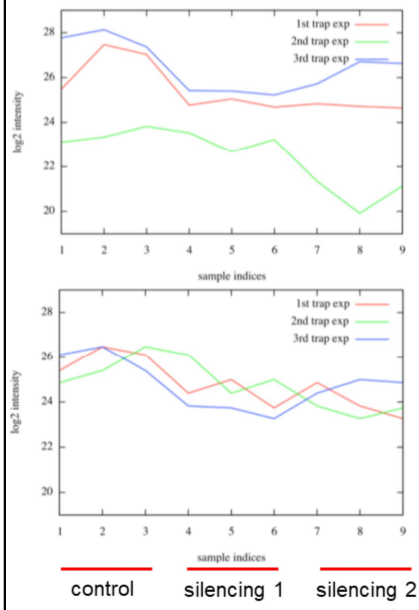
Imputation strategy – at least one valid data point

For proteins having at least one valid MS data point for knock-down samples, missing data points were generated from the valid data point(s) based on the local least squares (LLS) imputation method (see slide 23-25 of V3).

Subsequent to data imputation, we **log2-transformed** the ratio between siRNA and control siRNA samples,
and applied **protein-based quantile normalization** to homogenize the abundance distributions of each protein with respect to statistical properties.

If one valid data point is available, we felt that the additional imputed data points should be generated in the vicinity of this data point and not at the bottom of the distribution.

Protein-based quantile normalization



Intensity profiles for TRAP beta subunit across the 3 experiments before (top) and after (bottom) protein-based quantile normalisation.

Horizontal axis : sample IDs.
- 1 to 3 - control,
- 4 to 6 – SSR2 silencing by 1st siRNA,
- 7 to 9 – SSR2 silencing by 2nd siRNA.

Aim of protein-based quantile normalization: remove the systematic variation among 3 iterations of TRAP silencing experiment.

QN ranks the raw data, computes the averages, and replaces the original values by the ranked averages.

V3 Processing of Biological Data SS 2020 34

Shown are the log2-transformed data points for the beta-unit of TRAP.

The task was how to homogenize the data from 3 independent experiments.

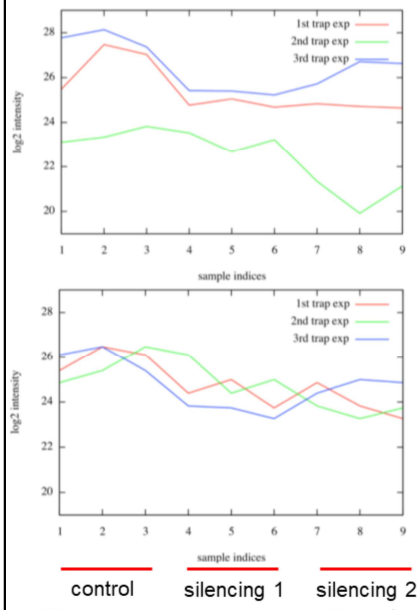
Typically, one applies quantile normalization on the full data set of all genes (proteins).

However, this did not work here. When clustering the data after normalization, data that should belong to each other was not clustered together.

Therefore, we used quantile normalization for the data points of each single protein.

As will be shown later, this worked quite well.

Protein-based quantile normalization



The variation left after normalisation reflects the biological variation between samples.

In the top panel, SSR2 levels of the controls (indices 1-3) are higher than both siRNAs in experiment 2 (red) and higher than the first siRNA in experiment 1 (blue).
In the third experiment (green), the second siRNA (indices 7-9) induces lower levels than in the controls and the first siRNA.

The same conclusions can be drawn from the bottom panel. The benefit of the normalized values here is that the blue, red, and green distributions contain identical values.

Thus, one can now apply standard statistical tests to identify the significant differences.

V3 Processing of Biological Data SS 2020 35

No comments.

Detection of differential abundance

Abundance in each siRNA knock-down was individually compared against control.

Proteins with an FDR-adjusted *p*-value (i.e. *q*-value) of below 5% were considered significantly affected by the siRNA knock-down.

Then, we intersected the results from the two unpaired *t*-tests for the 2 siRNAs.

This means that the abundance of all reported candidates had to be statistically significantly affected **in both siRNA silencing experiments**.

The available experimental data was quite difficult to handle.

We only had very few experimental data points. Also, a considerable portion of the data points were imputed. Furthermore, the trends found in the two silencing experiments were not always consistent.

Thus, we were quite strict in the statistical analysis.

We kept only those proteins that are significantly deregulated when comparing the first silencing siRNA against control AND when comparing the second silencing siRNA in the SAME DIRECTION.

In this way, we may have omitted some actual Sec61 or Trap clients, but we wanted to be rather conservative.

Comment by reviewer

6. *Statistical analysis of the data:*
On page 29 you describe imputation of data points.

Did you do a statistical analysis if the number of data points is sufficient that this imputation will not change results?

One reviewer of our manuscript challenged us to check how strongly data imputation affected the obtained results.

Validation of imputation method

Our reply: We assumed that … missing values … stem from "the bottom" of the distribution and belong to low abundance proteins that were not detected by the mass spectrometry instrument.

We tested to what extent the data imputation may affect the differential abundance analysis. ... The first Sec61 silencing experiment was selected for the validation… We selected only those proteins that have a "complete" dataset, i.e. none of out of nine entries was missing… This was the case for 5715 out of 6960 proteins….

To generate a synthetic dataset for missing data, we randomly removed 10% of the (known) data points from the lower tail of the distribution ...

For two different thresholds (5th and 10th percentile of the overall distribution), we repeated the removal 100 times. Therefore, in total, we generated 200 new datasets with artificially generated "missing" data.

Therefore, we did a test on proteins having a full data set with nine out of nine abundance values and randomly removed 10% of all their data points with low values.

Validation of imputation method

Subsequently, these “missing” data points were imputed.

Then, a differential protein abundance analysis was carried out on the imputed and the original data.

Finally, we compared the results of the differential analysis of the imputed and original data to validate the reliability of the imputation method.

For this, using the results of the previous steps, the significantly affected proteins were either labelled as 1 (positively affected) or as -1 (negatively affected) while the unaffected proteins were labelled 0.

Afterwards, we computed the Pearson correlation coefficient between the results of the original data and of the imputed data.

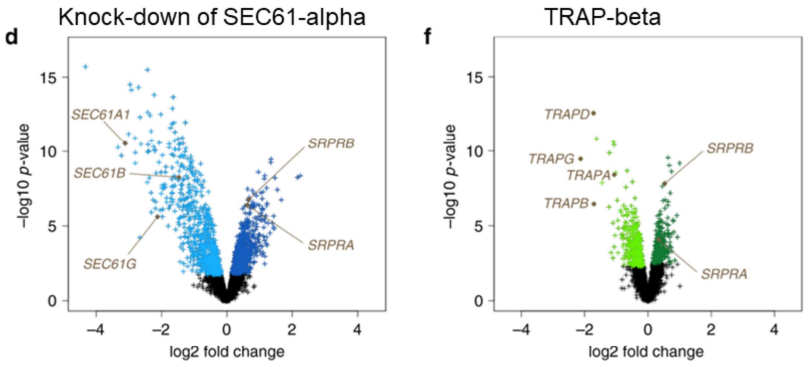
The overall correlation coefficients for the 5th and 10th percentile thresholds are 0.975 ± 0.018 and 0.927 ± 0.020, respectively.

The test showed that the results obtained for the imputed data were strongly correlated with the results for the complete data.

Of course, imputing data still introduced some artefacts in the analysis.

But this check shows that the magnitude of those artefacts appears tolerable.

Volcano plot: differential protein abundance



Differentially affected proteins were characterized by the mean difference of their intensities plotted against the respective permutation false discovery rate-adjusted *p*-values in volcano plots.

The results for a single siRNA are shown in each case (*SEC61A1*-UTR siRNA, *TRAPB* siRNA).

V3 Processing of Biological Data SS 2020 40

The left panel shows which proteins are downregulated if the alpha-subunit of Sec61 is silenced.

Of course, the subunits of Sec61 itself are downregulated as expected.

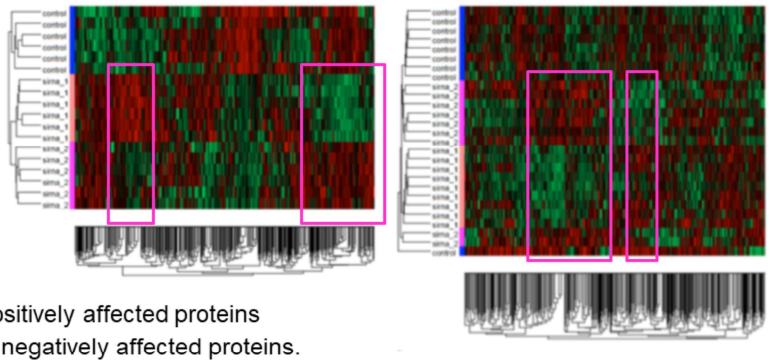
On the other hand, the cell upregulates the 2 subunits of the SRP receptor (SRPRB and SRPRA) that usually guide nascent peptide chains from the ribosome to the translocon because the cell senses that something is wrong with protein translocation. So this is a rescue mechanism. In fact, the cell actually upregulates a number of other proteins as well.

The right panel shows which proteins are downregulated if the beta-subunit of Trap is silenced. Overall, these are fewer proteins than in the left panel. This makes sense because about 1/3 of all cellular proteins need to pass Sec61, but only a portion of them also need Trap.

Up- / down-regulation

Heat maps visualize clusters of proteins that were

- significantly **upregulated** following treatment with both siRNAs directed against either *SEC61A1* (left) or *TRAPB* (right) mRNA or with non-targeting (control) siRNA, or that were
- significantly **downregulated**, or that
- represent **variations** between siRNAs.



Red : positively affected proteins
Green : negatively affected proteins.

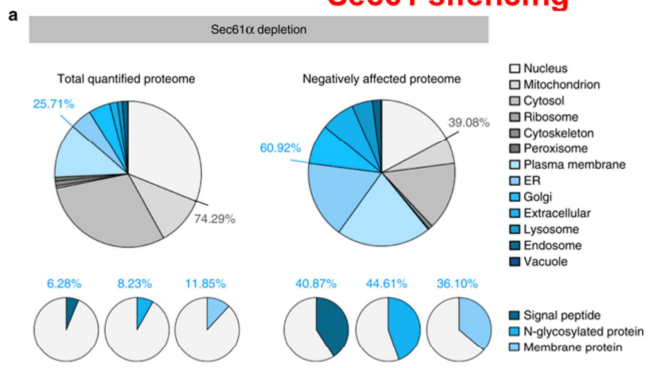
V3 Processing of Biological Data SS 2020 41

The heatmap shows 2 points, one is good, one is problematic.

The good point is that for Sec61alpha silencing (left panel), clustering by control/siRNA1/siRNA2 worked perfectly. Also for Trap silencing (right panel), clustering worked quite well. Only the bottom 3 data rows are clustered away from the other experiments.

The problematic point is that the results for the two silencing siRNAs are sometimes inconsistent. I have enclosed some of the problematic regions with pink boxes.

Annotation of differentially abundant proteins after Sec61 silencing



Validation of Sec61 clients based on Gene Ontology enrichment factors.

Protein annotations of signal peptides, membrane location, and N-glycosylation in humans were extracted from UniProtKB, and used to determine the enrichment of Gene Ontology annotations among the secondarily affected proteins.
Summary of two Sec61 depletion experiments performed in triplicate.

Here, we annotate the identified downregulated proteins by their GO localization.

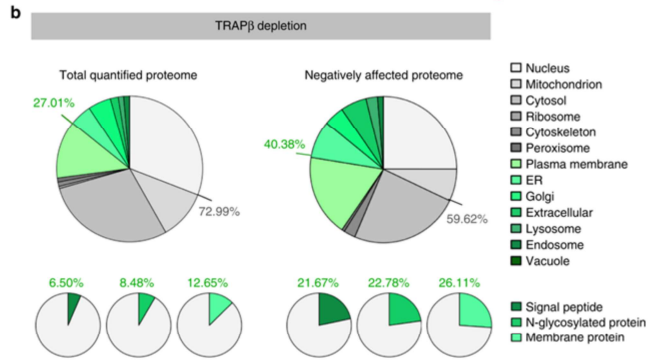
Blue colored compartments in the „cake“ belong to the secretory pathway and to membrane compartments.

Compared to all proteins identified by MS (left), the downregulated proteins are more than 2-fold enriched in these compartments as expected.

39% of the hits localize to other compartments. These proteins are not expected to be Sec61 clients themselves. Their downregulation may either be a compensatory biological effect or simply be due to experimental noise.

In the lower line, we analyzed how many of the proteins have signal peptides, are glycosylated or are membrane proteins. All these properties are strongly upregulated as expected.

Annotation of differentially abundant proteins after TRAP silencing



Validation of TRAP clients based on Gene Ontology enrichment factors.

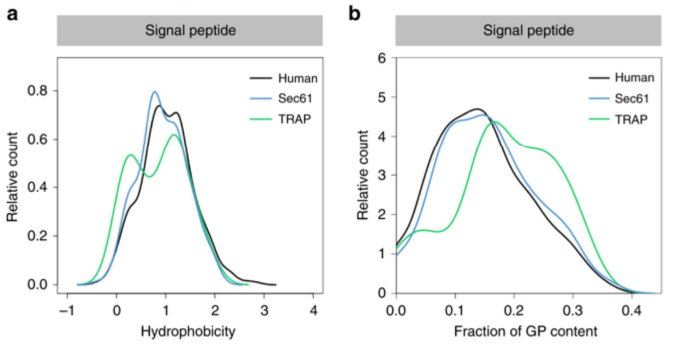
Summary of three TRAP depletion experiments performed in triplicate.

→ clear enrichment of green fraction (ER targeted organelles)

V3 Processing of Biological Data SS 2020 43

This is the same analysis for the downregulated proteins after Trap silencing.
Now, the enrichment of relevant compartments is only about 1,5-fold.
Also, fewer relevant features are found in the lower row.

Physicochemical properties of TRAP clients



The signal peptides of TRAP clients are **less hydrophobic** and have a **higher Gly/Pro content** than Sec61 clients and the full proteome.

Physicochemical properties of TRAP clients with signal peptide (SP).

Hydrophobicity score (**a**) and glycine/proline (GP) content (**b**) of SP sequences. Hydrophobicity score was calculated as the averaged hydrophobicity of its amino acids according to the Kyte-Doolittle propensity scale. GP content was calculated as the total fraction of glycine and proline in the respective sequence.

Here, we tried to identify whether the signal peptides of TRAP clients differ from the background of cellular proteins.

Indeed, we found that their signal peptides are less hydrophobic (left panel) and contain more Glycine and Proline residues – which can be expected to weaken their helical propensity.

Therefore, one can speculate that these nascent peptide chains cannot push the Sec61 pore open by themselves and need to be aided by the adjacent Trap complexes to open the Sec61 pore.

Summary

Mass spectrometry is the method of choice to characterize the cellular proteome.

The good point about MS is the high sensitivity and resolution: one can easily detect posttranslational modifications.

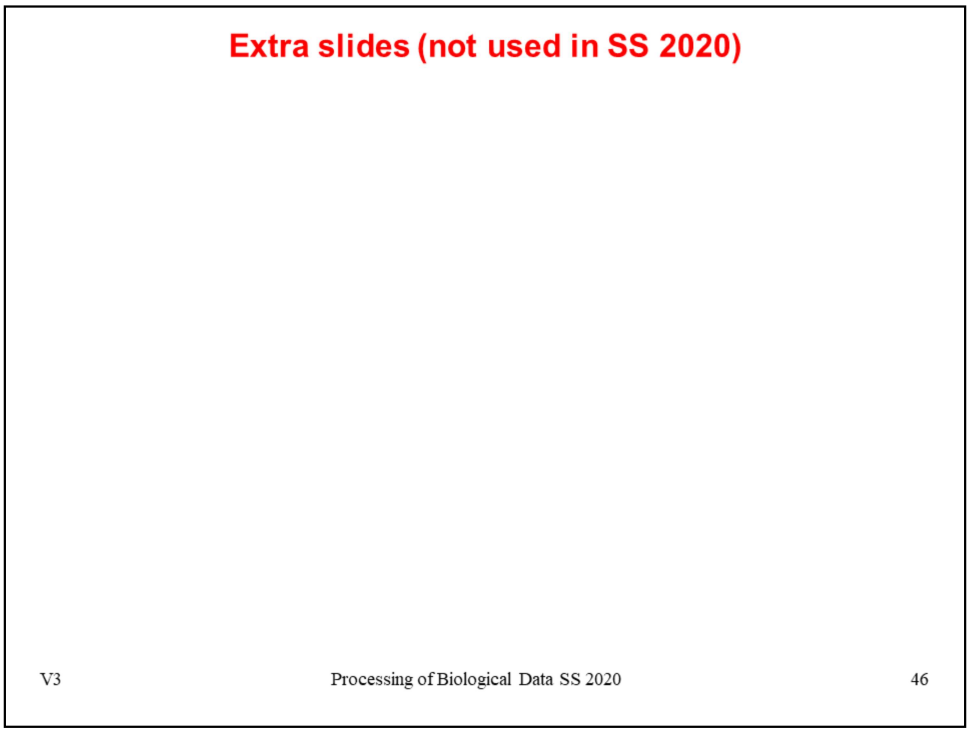
However, MS instruments are very expensive to buy and to operate → usually we have much fewer datasets available than from transcriptomics experiments.

In terms of impact, proteomics analysis is 5 - 20 years behind transcriptomics analysis.

Dealing with missing data points is a big challenge in proteomics.

Although mRNA copy numbers and protein copy numbers are generally correlated somehow, there are often surprises when synthesis rates and/or half-lives are not matching to each other.

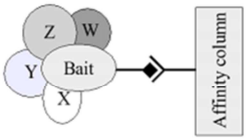
No comments.


Extra slides (not used in SS 2020)
V3
Processing of Biological Data SS 2020
46

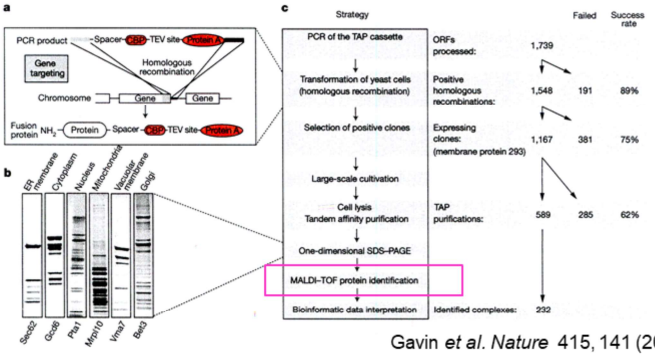
No comments.

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)

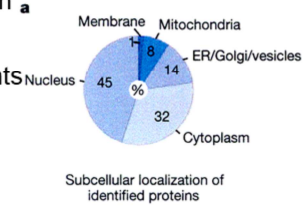
This is an example of MS in detecting which proteins belong to protein complexes.

The paper by Anne Gavin et al. is a classic paper on the application of the TAP-MS method (https://www.ncbi.nlm.nih.gov/pubmed/11805826) that has been cited more than 5500 times.

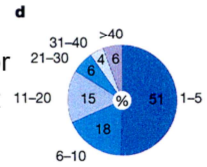
TAP analysis of yeast PP complexes

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%)



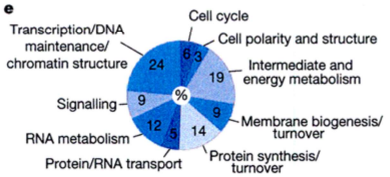
Subcellular localization of identified proteins



Number of proteins per complex

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

(e) Complexes are involved in practically all cellular processes



Distribution of complexes according to function

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

This is slide illustrates what proteins belong to the 232 identified complexes of yeast.

Panel a shows that the proteins belong to different compartments.

Panel d shows the size of the complexes (# of proteins)

Panel e shows the biological processes carried out by the identified proteins.

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.

Let α and β be the rate of missing values and the MNAR ratio, respectively.

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

This study (https://pubs.acs.org/doi/10.1021/acs.jproteome.5b00981) investigated the effect of assuming 3 different models for missing values.

MCAR, MAR and MNAR are standard models in data science (see https://en.wikipedia.org/wiki/Missing_data).

Simulation benchmark

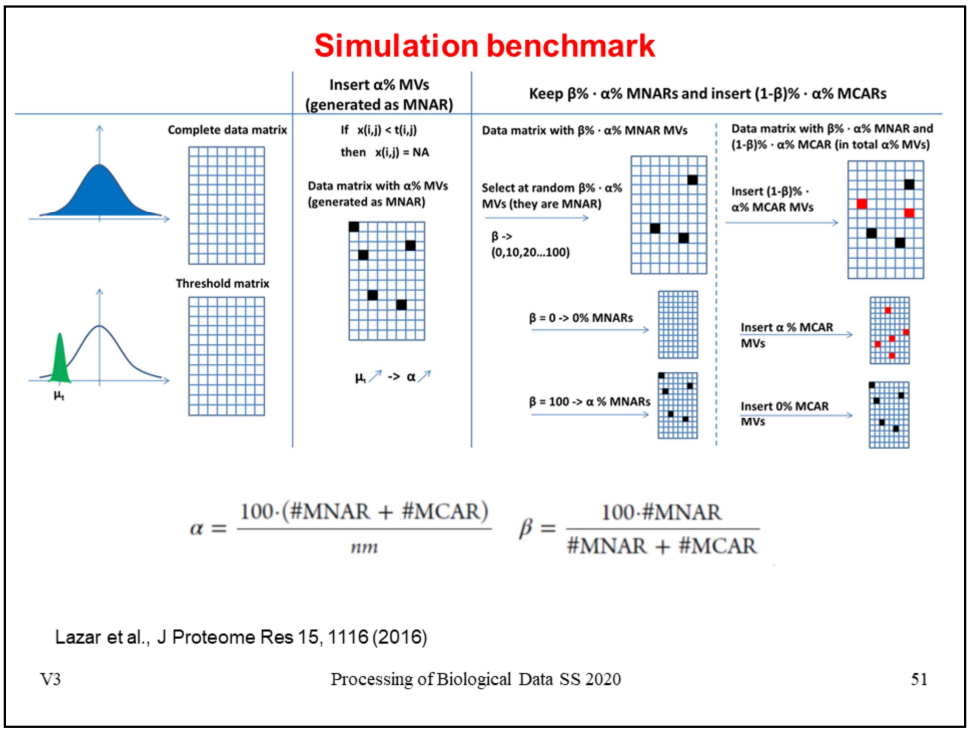
Use real data (Super-SILAC and label-free quantification) on human primary tumor-derived xenograph proteomes for the two major histological subtypes of nonsmall cell lung cancer : adenocarcinoma and squamous cell carcinoma.

MNAR values: one randomly generates a **threshold matrix** T from a Gaussian distribution with parameters (μ_t = q, σ_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 βα · 100 determines if the abundance value is censored (success) or not (failure).
α and β are the rate of missing values and the MNAR ratio, respectively.

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)

This slide describes how random values were inserted into a real data set. The procedure will be explained again on the next slide.



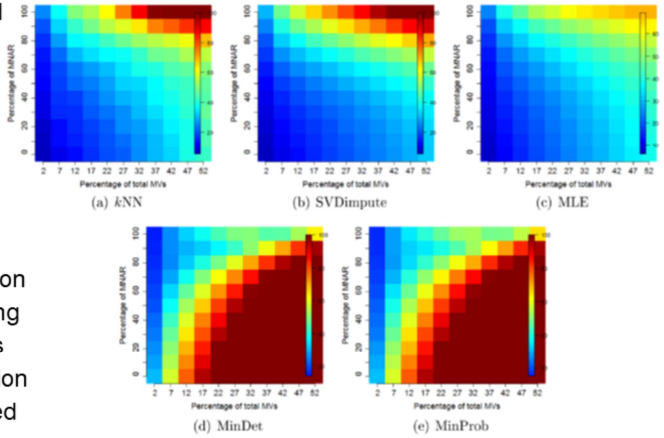
Schematic view upon the strategy used for the missing data generation. This strategy allows to control both for the total proportion of missing values generated as well as for the proportion of missing values, which are MNAR and MCAR.

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



(a) *k*NN (b) SVDimpute (c) MLE

(d) MinDet (e) MinProb

RSR = RMSE / std.dev. MV: missing value

Blue: low RSR

Red: high RSR

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

RSR for the real quantitative data set; imputation is performed by considering: *k*NN (a), SVDimpute (b), MLE (c), MinDet (d), and MinProb (e).

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.

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

Different algorithms provide advantages for different frequencies of missing values.