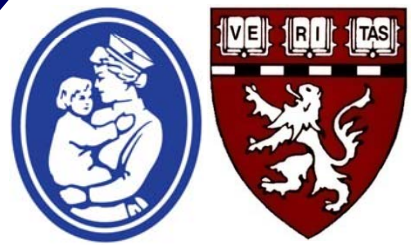
Joint BSPR/EBI Conference 2010
Proteomics: From Qualitative to Quantitative
Hinxton, July 13-15, 2010

new strategies for the identification of co-regulated proteins

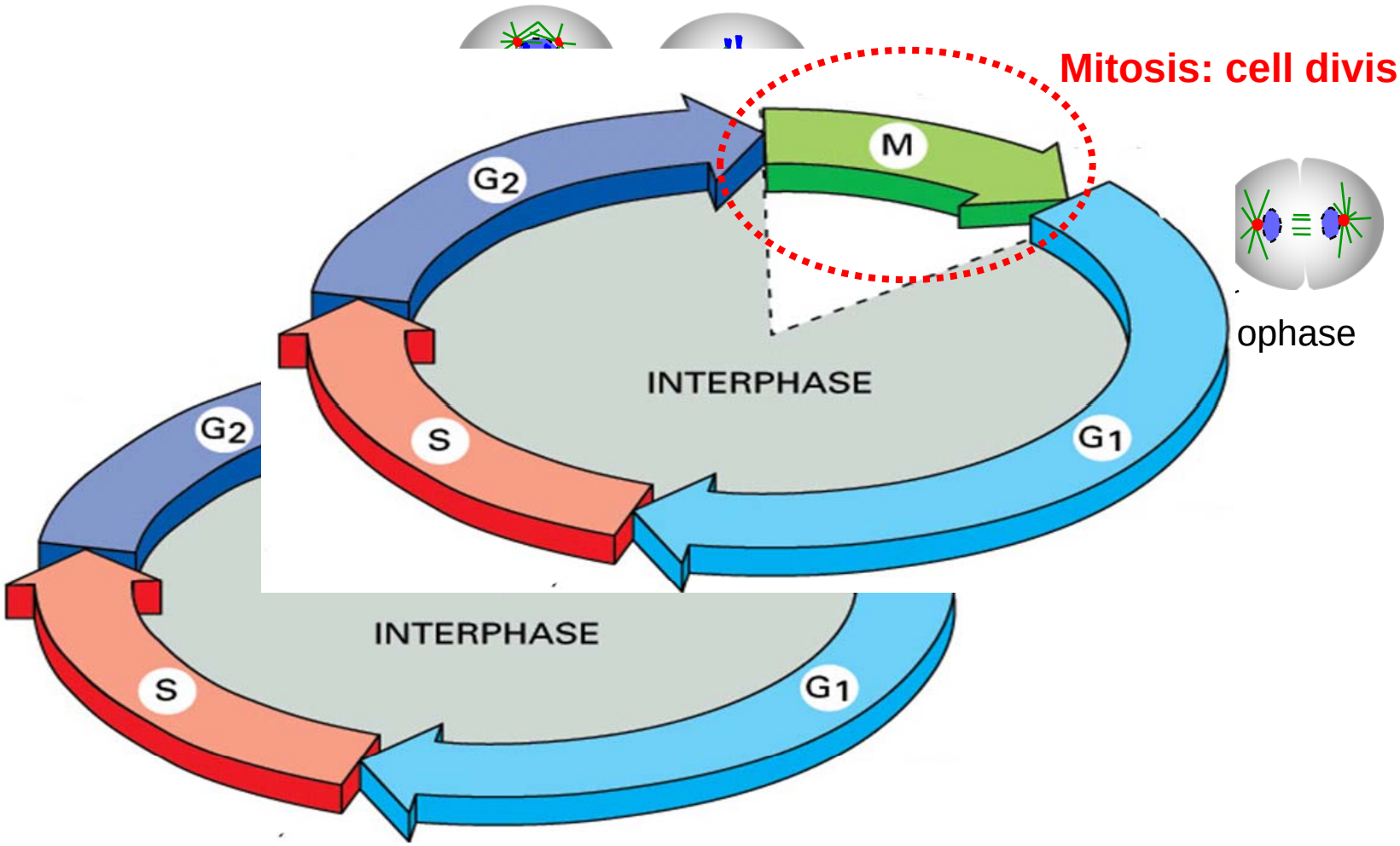
Hanno Steen, Ph.D.

Department of Pathology

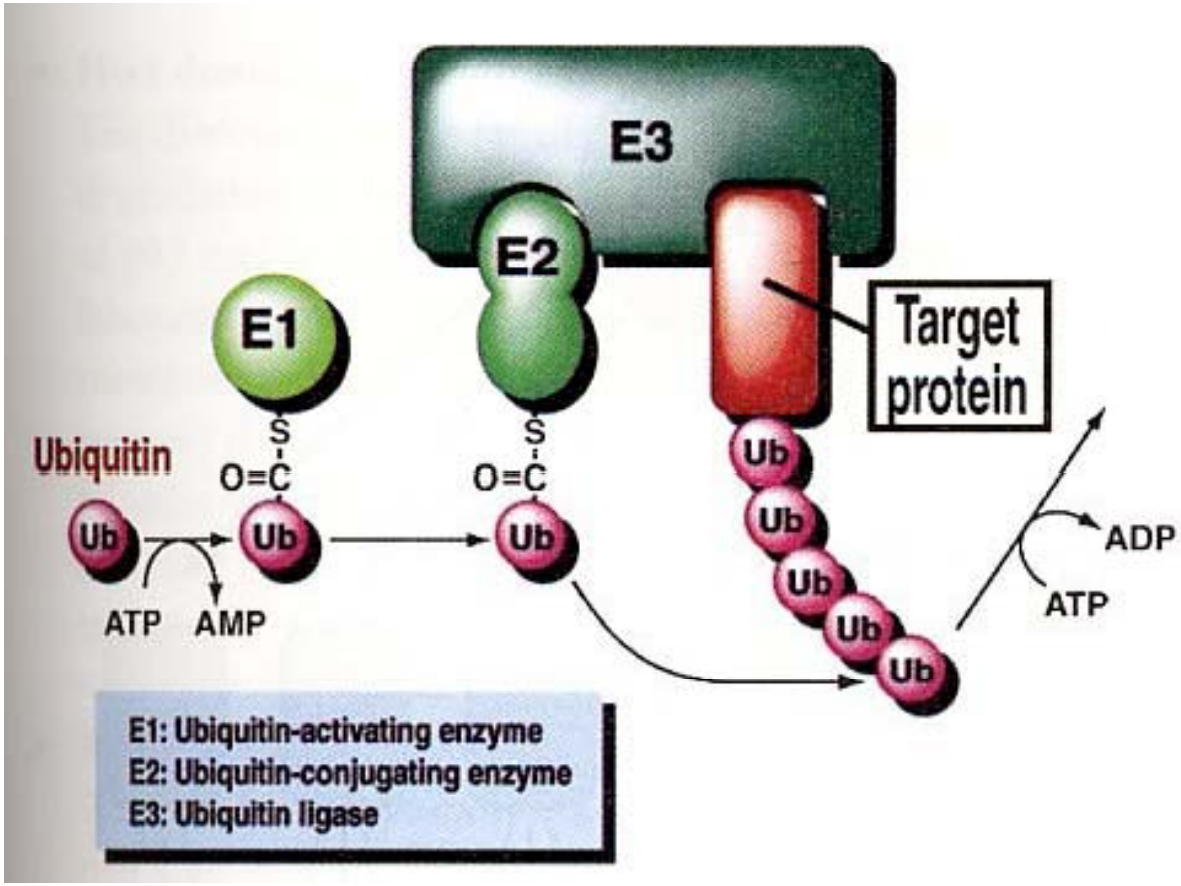
Harvard Medical School and Children's Hospital Boston



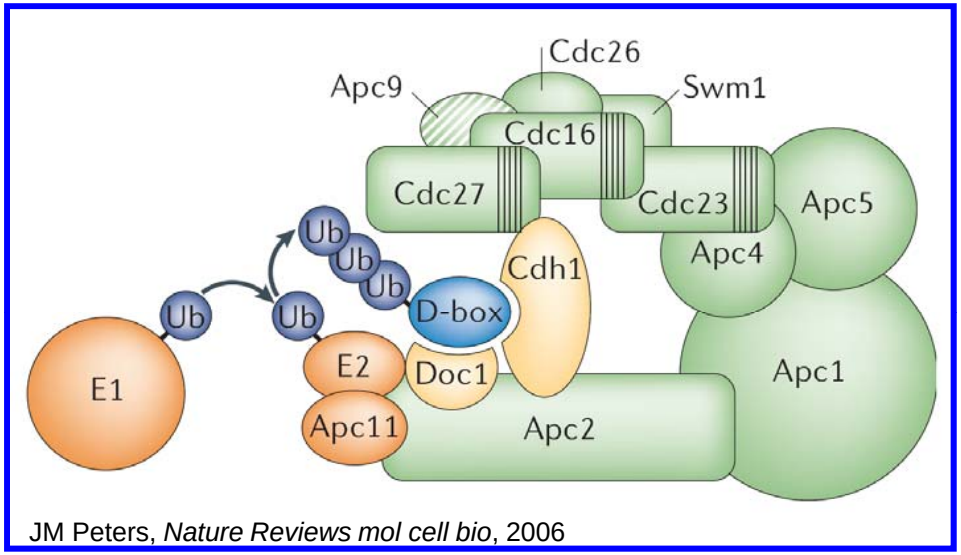
the cell cycle



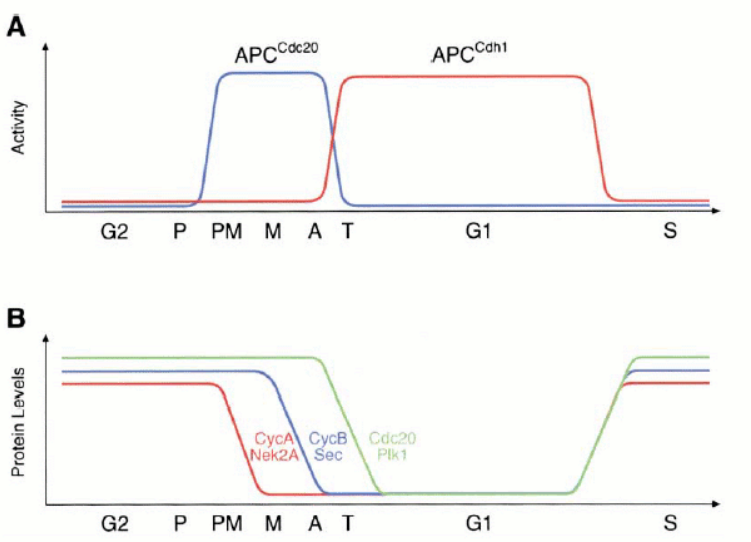
the ubiquitination/degradation pathway



APC – an E3-ligase
(anaphase promoting complex)



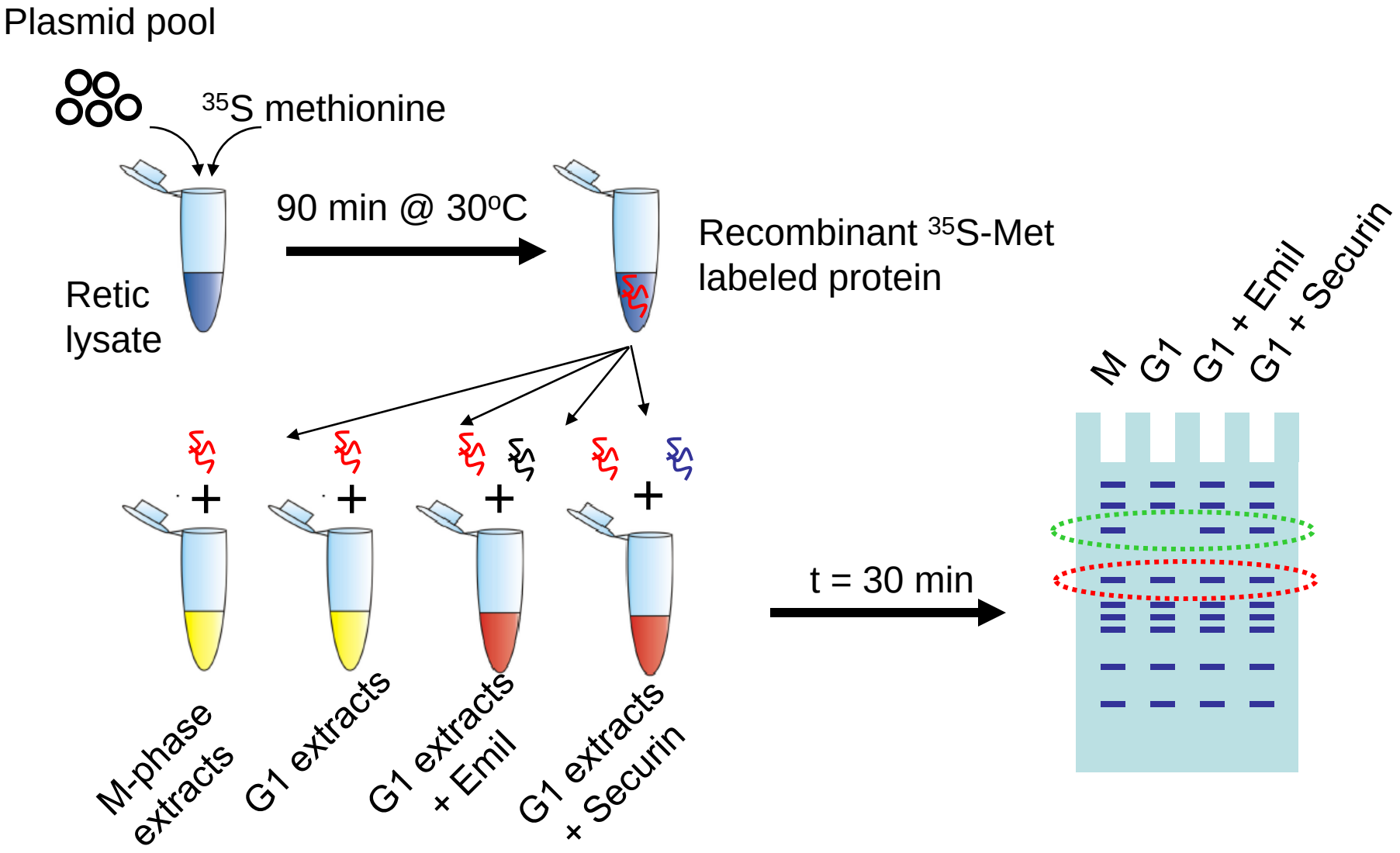
JM Peters, *Nature Reviews mol cell bio*, 2006



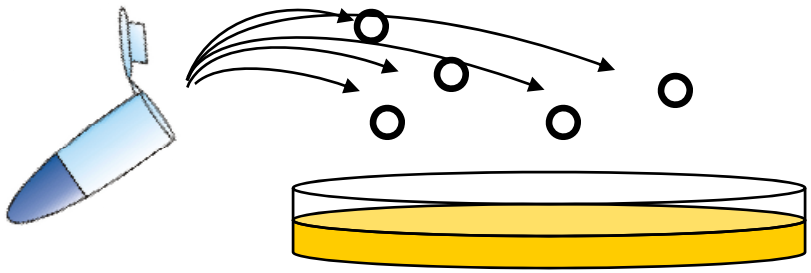
JM Peters, *Mol Cell,* 2002

- **Multi-subunit E3 ubiquitin ligase**
- **Initiates anaphase and mitotic exit and shows activity in G1**
- **Ensures specific and orderly degradation of numerous proteins….**
- ***…most of them are important cell cycle regulators!***

how to find APC substrates?

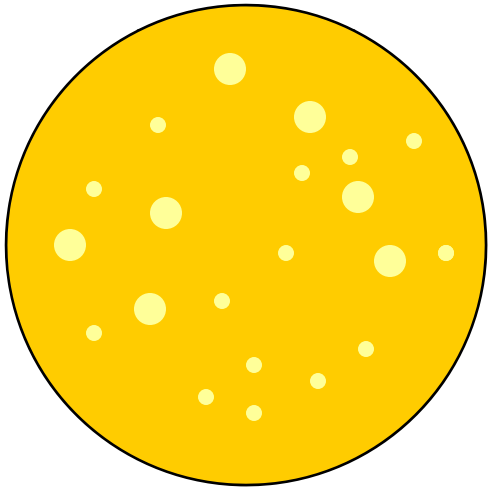


how to find APC substrates? (cont'd)

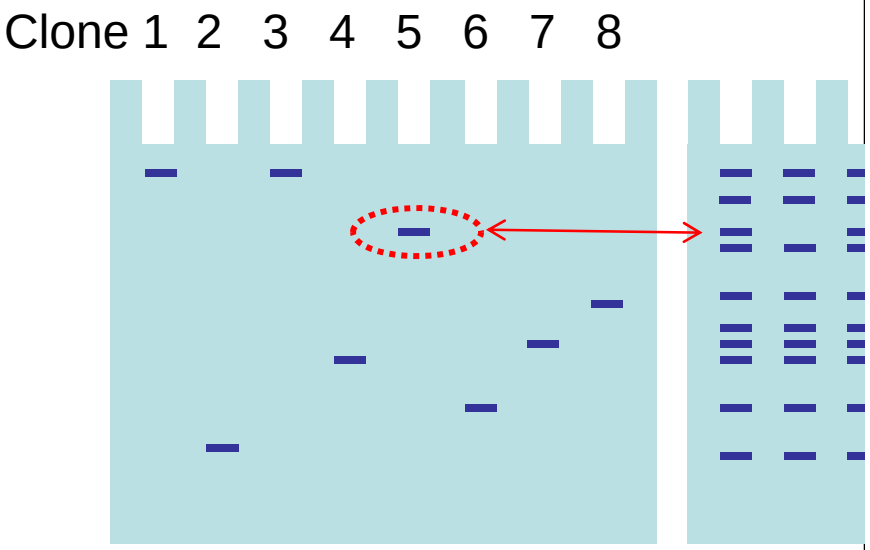


i) Plate the plasmid pool

ii) Let them grow



vii) Test clones
for degradation



iii) Pick colonies
iv) Grow them up
v) Prepare plasmid
vi) *In vitro* translate/transcribe plasmid

mass spectrometrist's view on the classic degradation assay

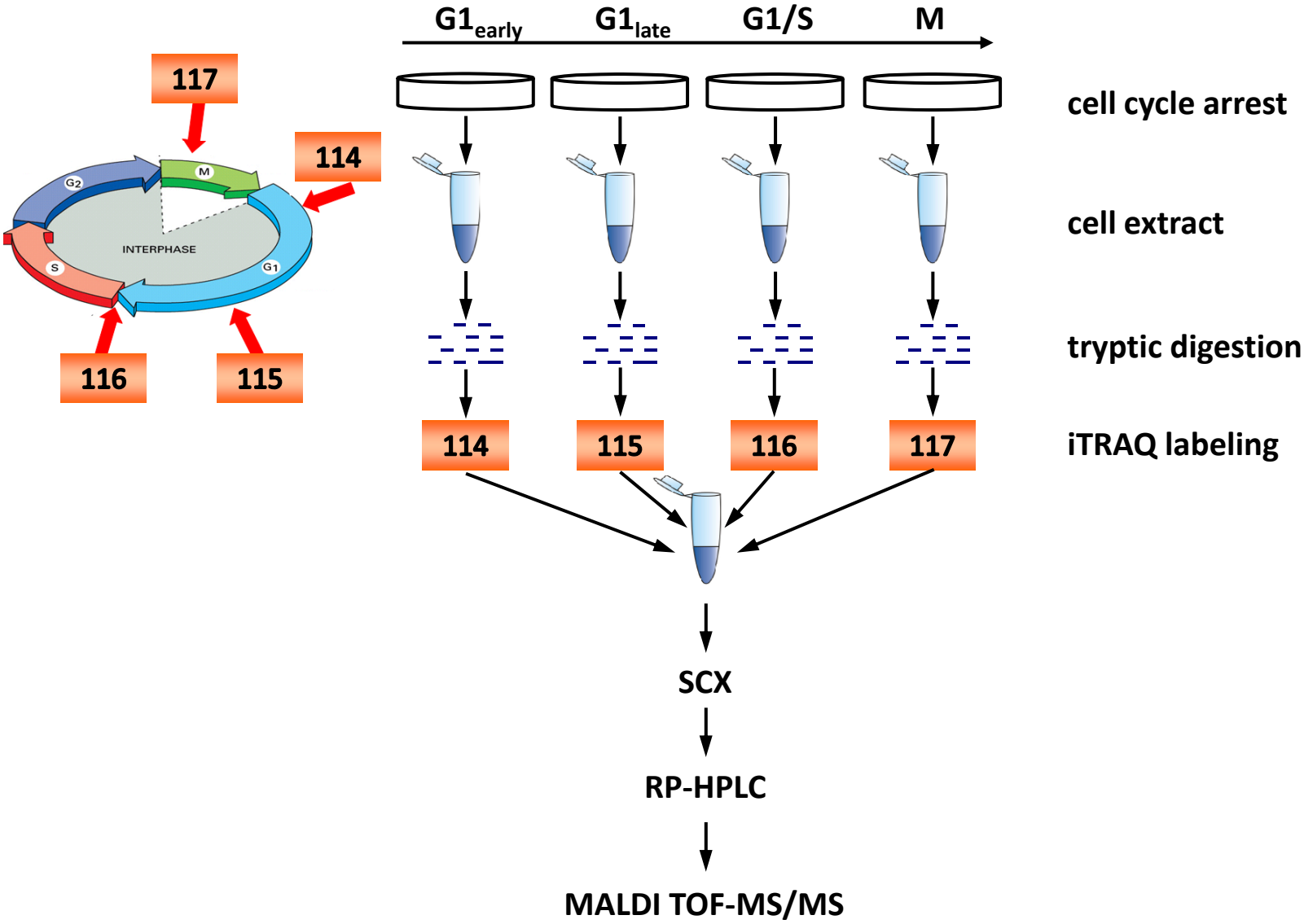
Far too tedious!

There must be better ways…

…and mass spectrometry might be one…

Hypothesis: Quantitative proteomics and hypothesis-driven data analysis can be used to identify targets of the APC/E3 ubiquitin ligases?!

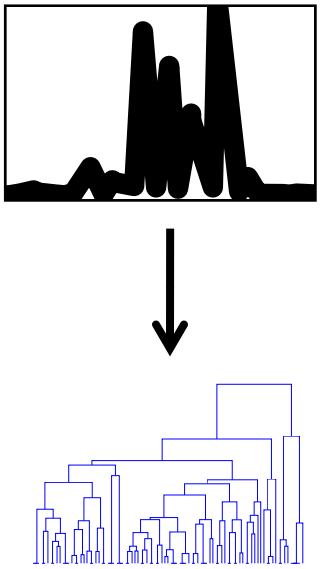
iTRAQ meets cell cycle – the workflow



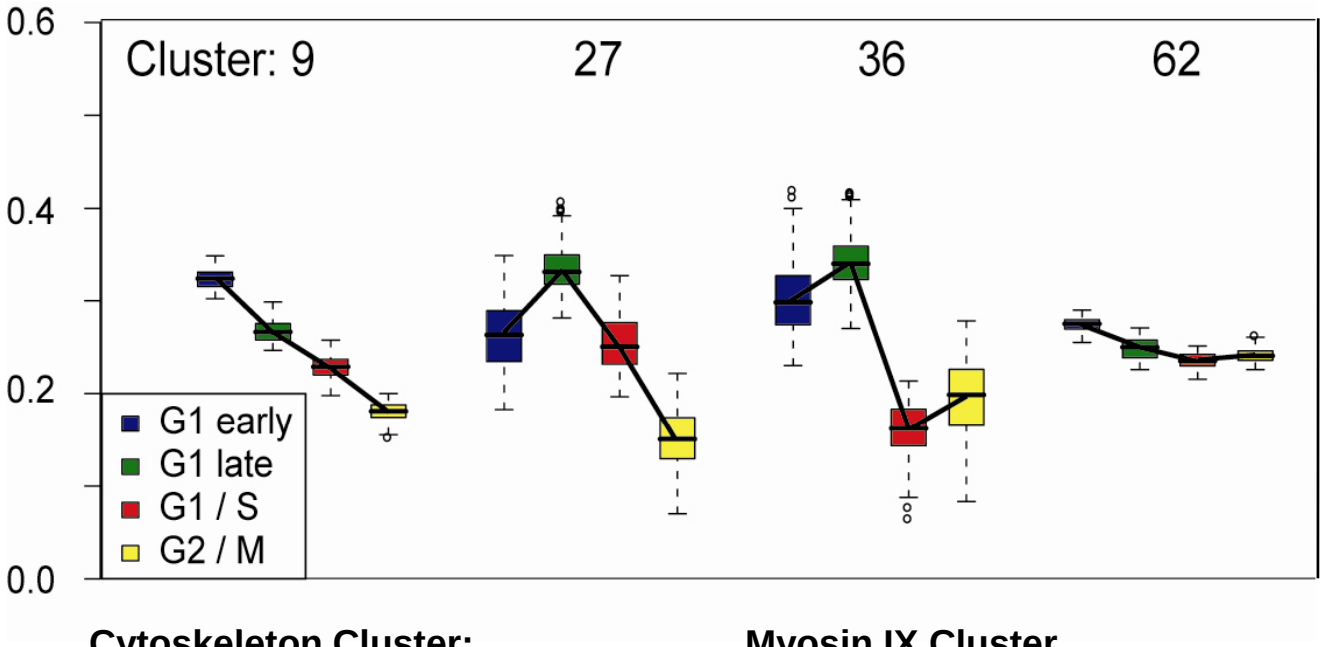
the data mining workflow

1. **Extract iTRAQ reporter ion intensities.**

2. **Cluster all *peptides* according to their quantitative profiles to group all peptides with similar temporal variations.**



the clusters (cont'd)



Cytoskeleton Cluster:
- lamin
- filamin
- plectin

Myosin IX Cluster

Ki-67 Cluster

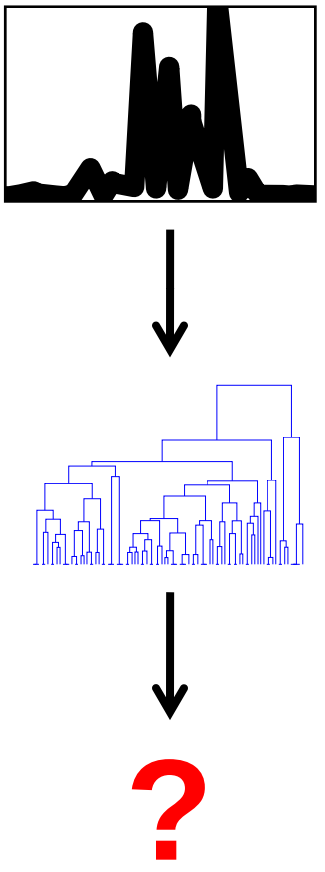
Metabolic Enzyme Cluster:
- pyruvate kinase 3

the data mining workflow

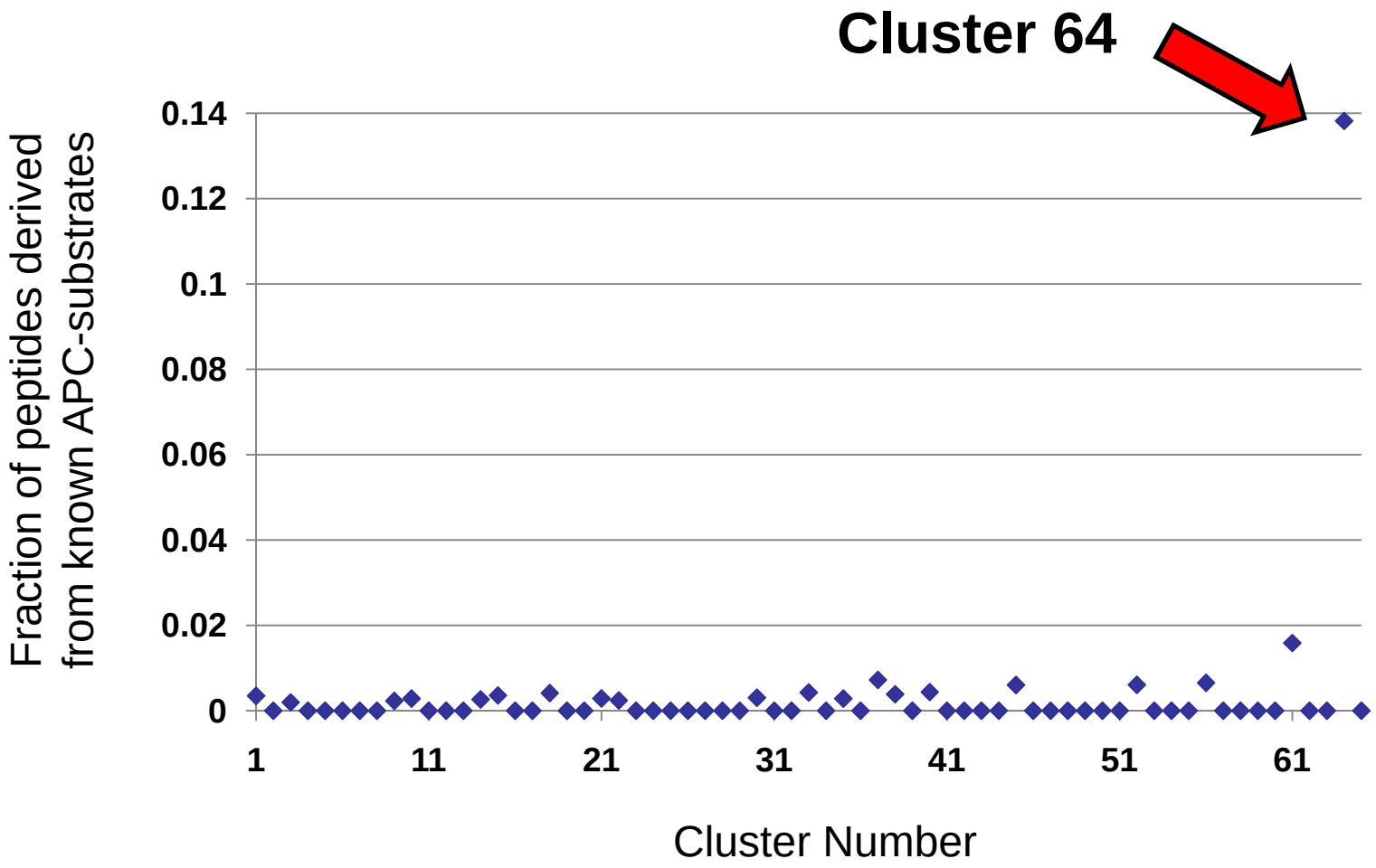
1. **Extract iTRAQ reporter ion intensities.**

2. **Cluster all *peptides* according to their quantitative profiles to group all peptides with similar temporal variations.**

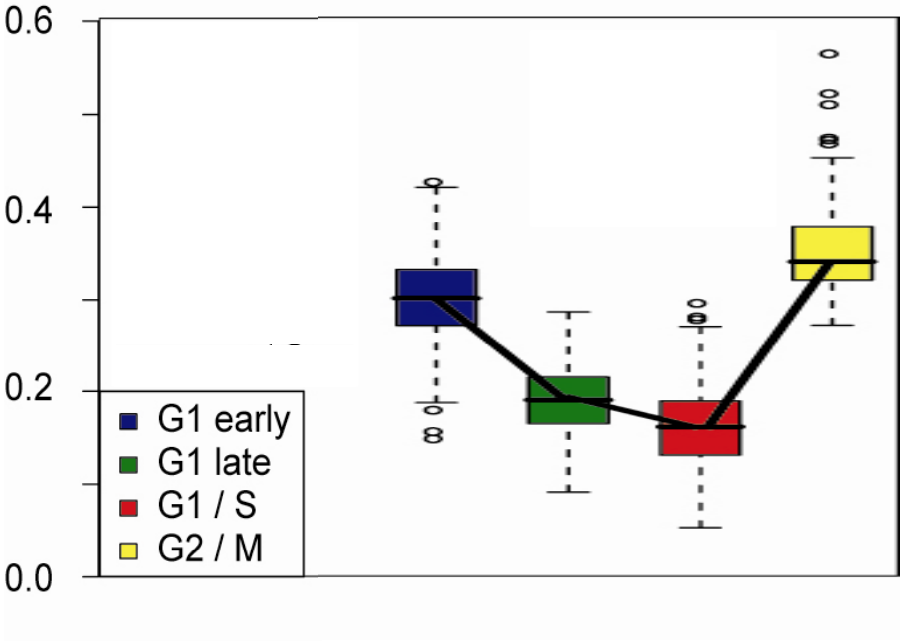
3. **Define clusters of interest.**



any known APC substrates?



cluster 64



Bottom line: Shows highest protein levels in prometaphase with subsequent decrease during G1.

14 % of all peptides are derived from known APC substrates.

11 out of 12 cyclin B1-derived peptides found in Cluster 64.

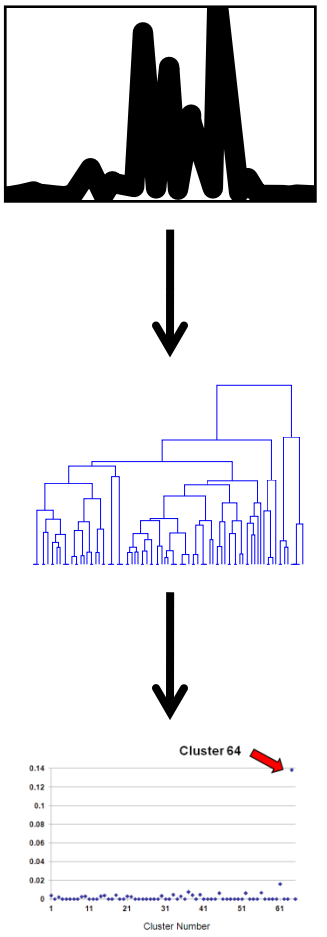
Cluster 64 is the cluster of interest!

the data mining workflow (cont'd)

1. **Extract iTRAQ reporter ion intensities.**

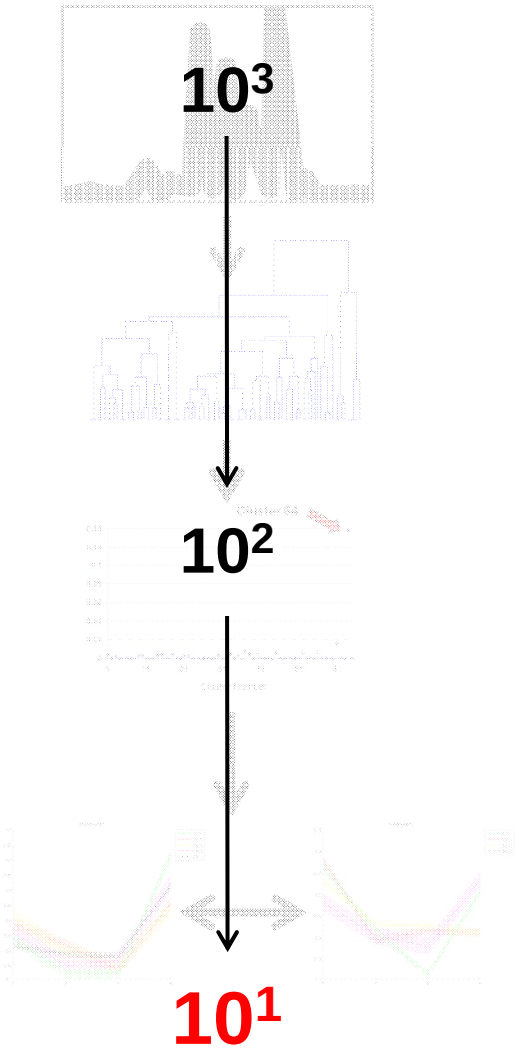
2. **Cluster all *peptides* according to their quantitative profiles to group all peptides with similar temporal variations.**

3. **Define clusters of interest.**

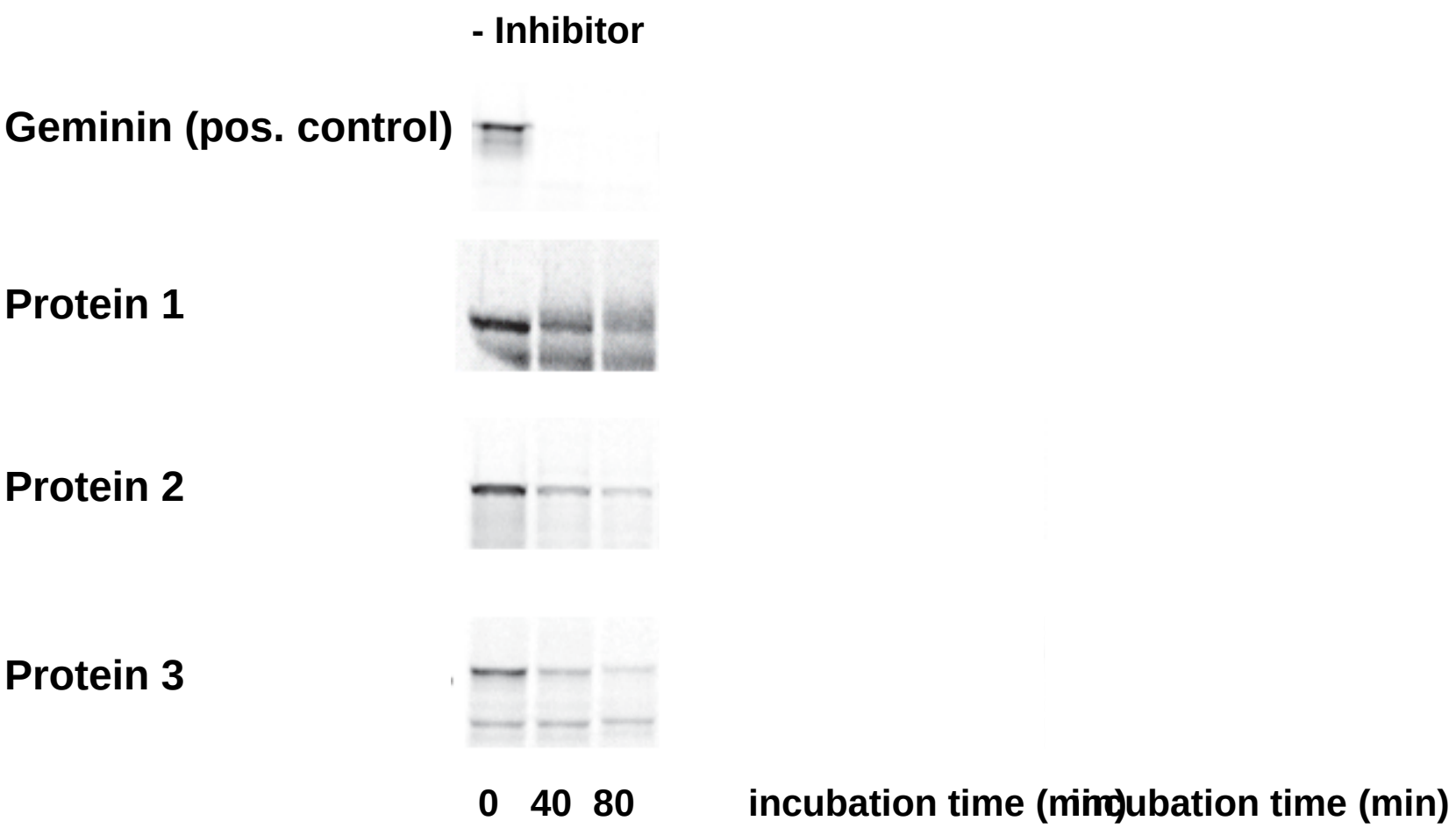


the data mining workflow (cont'd)

1. **Extract iTRAQ reporter ion intensities.**
2. **Cluster all *peptides* according to their quantitative profiles to group all peptides with similar temporal variations.**
3. **Define clusters of interest.**
4. **Identify *peptides* with highest similarity based on cyclin B1 model.**
5. **Calculate average profile for the *proteins*; determine statistical significance.**

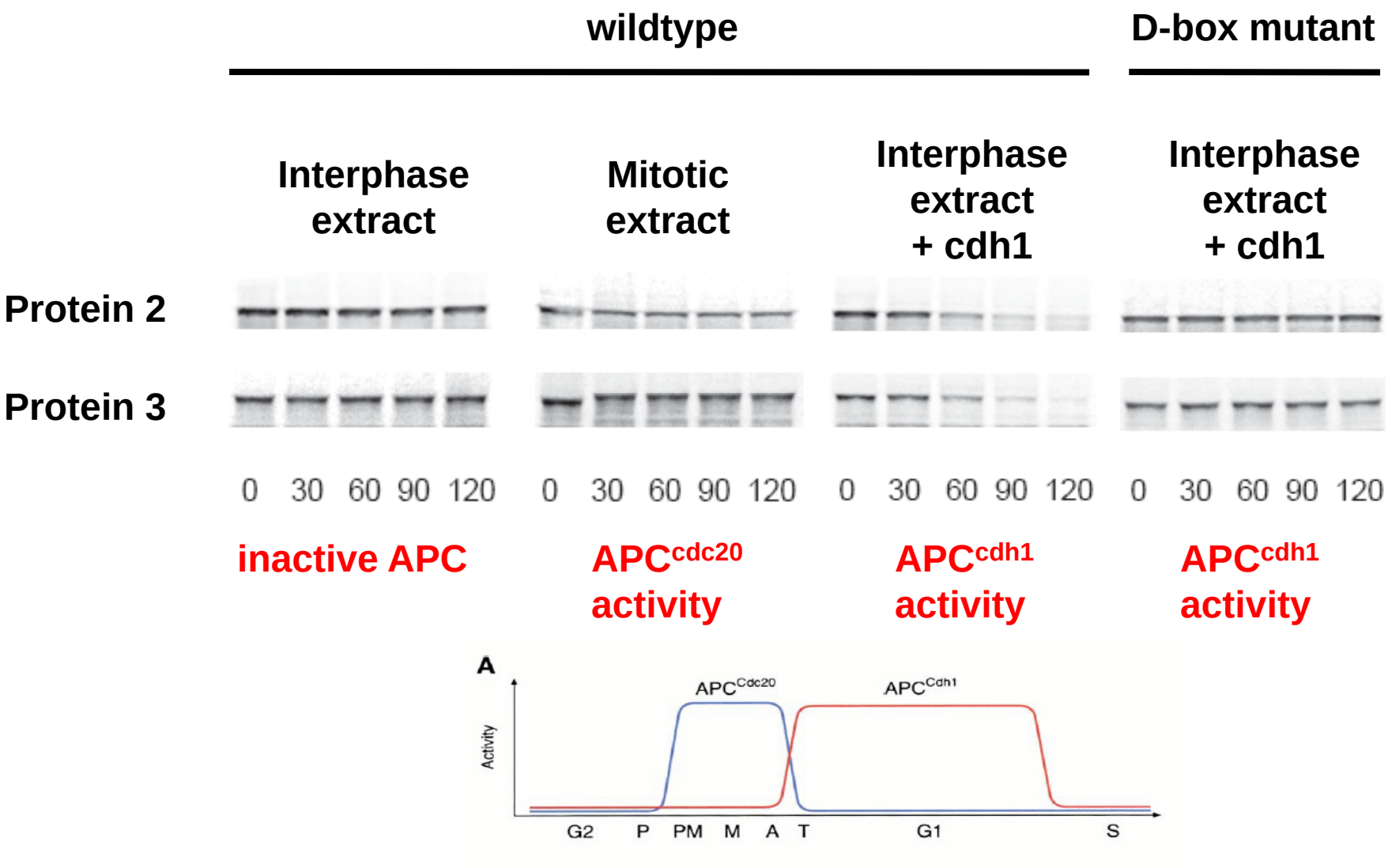


testing our shortlist



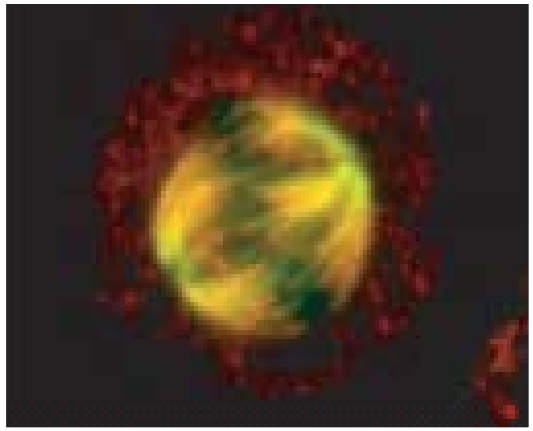
Incubation in G1 extracts of HeLa cells, i.e. APC^{cdh1} activity

APC^{cdc20} or APC^{cdh1} substrates?



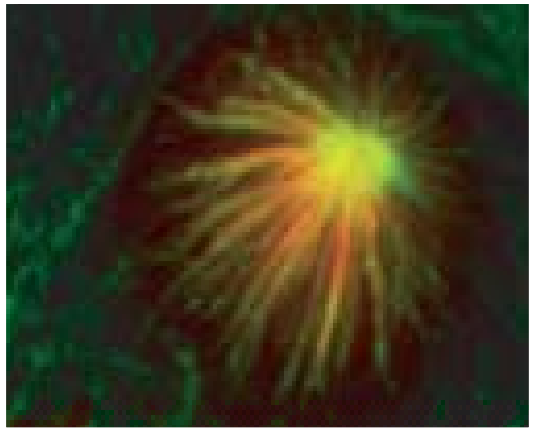
the bottom line…

Wildtype
Protein 2 is degraded



Bipolar spindle

D-box Mutant
Protein 2 is stable

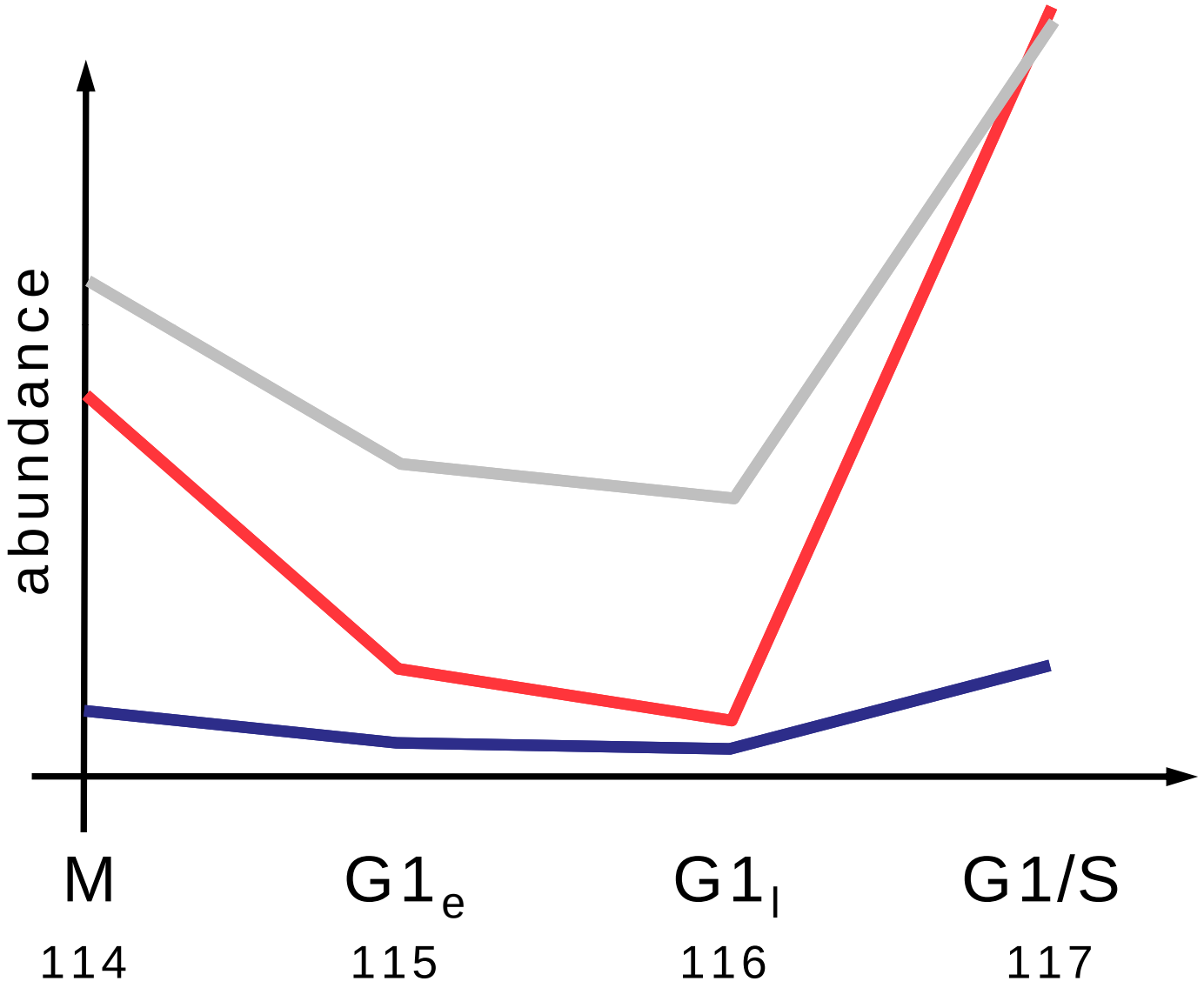


Monopolar spindle

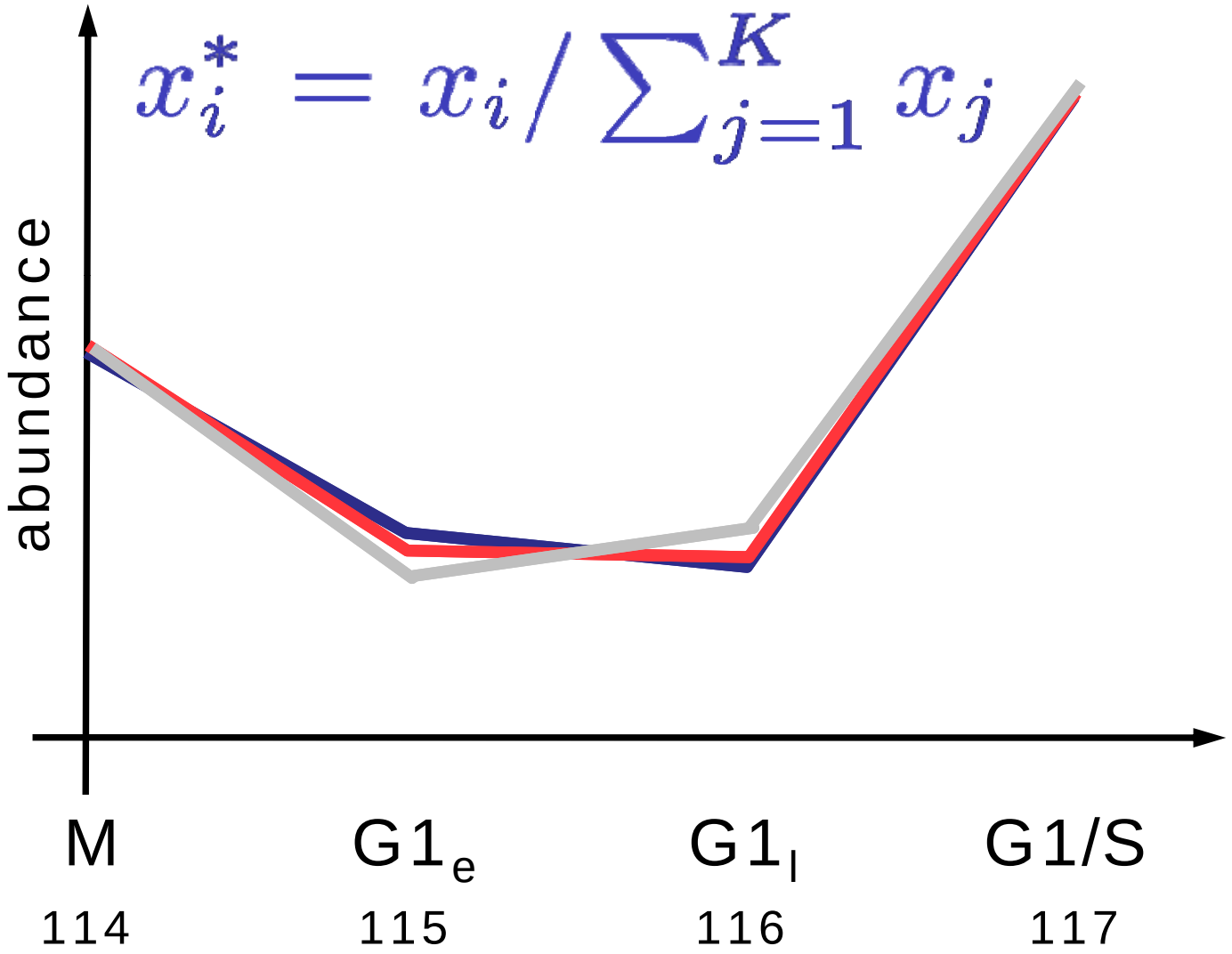
summarizing the prototype: an ms-based APC/C substrate screen

- Computational tools allows us to find the ‘needle in the haystack’.
- An *in silico* **50-fold enrichment of APC/C substrates** was achieved in the final shortlist.
- **2 novel APC substrates** were found and confirmed as APC^{cdh1} substrates.
- We confirmed our original hypothesis that quantitative proteomics and hypothesis-driven data analysis can be used to identify targets of the APC/E3 ubiquitin ligases!
- Moving forward: establishing appropriate statistical tools for this “**protein profile similarity screening**”.

the raw/absolute iTRAQ reporter ion intensities are not very useful...



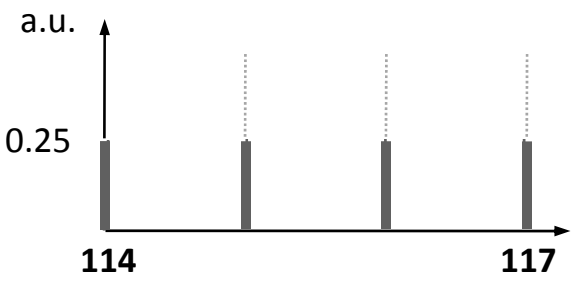
…thus, we have to normalize…

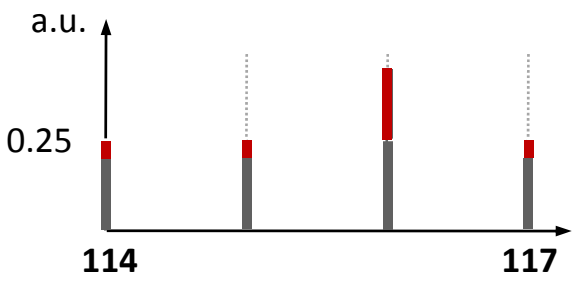


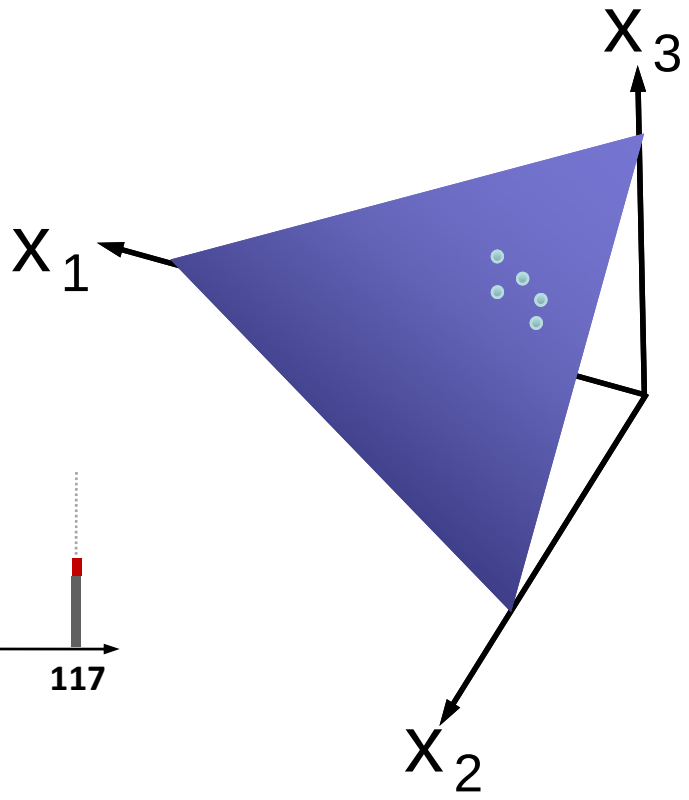
however: normalization is the root of all evil

- normalization causes variable dependency!

$$x_i^* = 1 - \sum_{j \neq i} x_j^*$$







however: normalization is the root of all evil (cont'd)

- simplicial data **do not** follow a normal distribution!

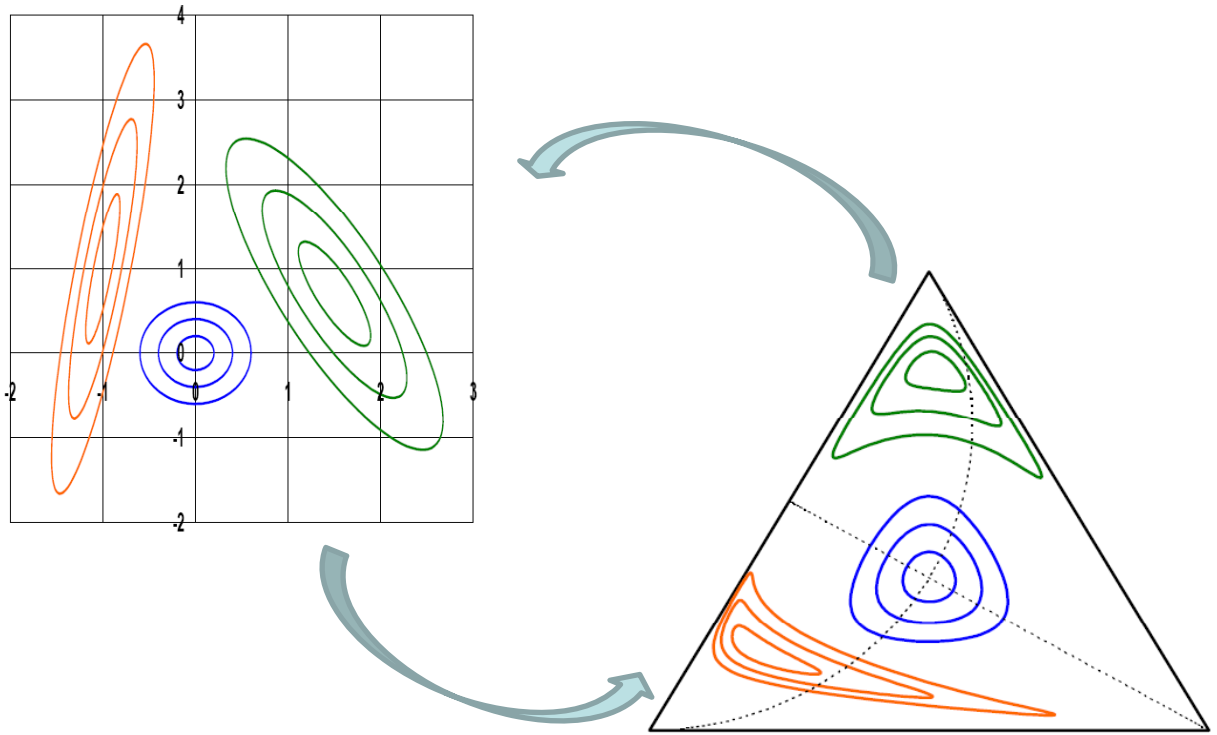


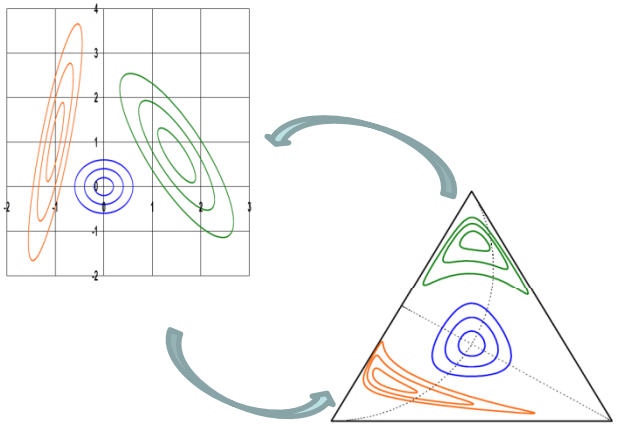
Figure adapted from Pawlowsky-Glahn *et al.*, Lecture Notes on Compositional Data Analysis, 2007.

however: normalization is the root of all evil (cont'd)

- simplicial data **do not** follow a normal distribution!

$$\mathcal{D}(\alpha_1, \ldots, \alpha_K) = \frac{\Gamma(\sum_i \alpha_i)}{\prod_i \Gamma(\alpha_i)} \prod_i (x_i^*)^{\alpha_i - 1}$$

Dirichlet distribution



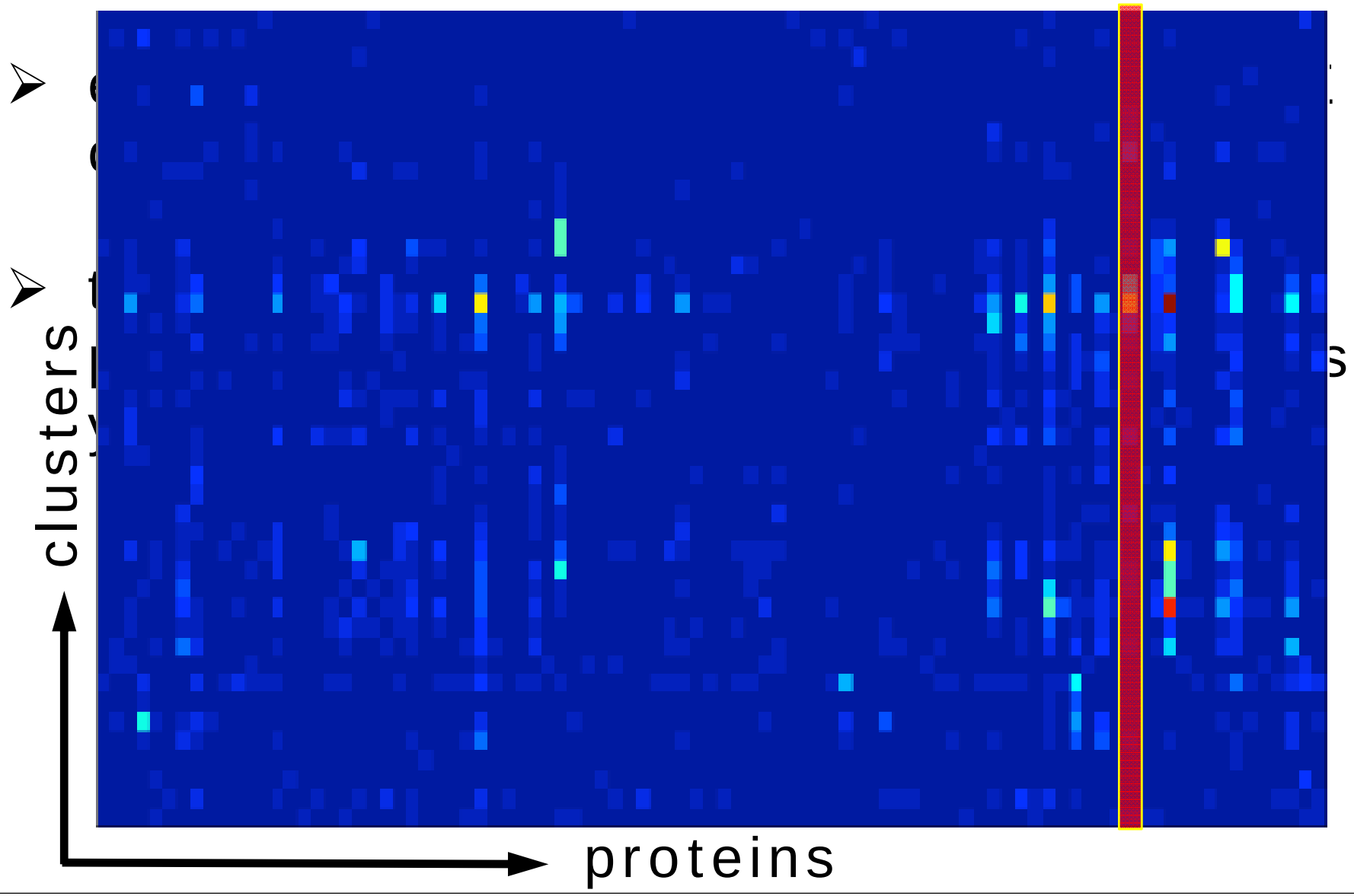
$$\lambda(\mathcal{X}, \mathcal{Y}) = \frac{L^{H_0}\left(\hat{\boldsymbol{\alpha}}^{\mathcal{X}\cup\mathcal{Y}} | \mathcal{X}, \mathcal{Y}\right)}{L^{H_1}\left(\hat{\boldsymbol{\alpha}}^{\mathcal{X}}, \hat{\boldsymbol{\alpha}}^{\mathcal{Y}} | \mathcal{X}, \mathcal{Y}\right)}$$

Dirichlet Likelihood Ratio Test

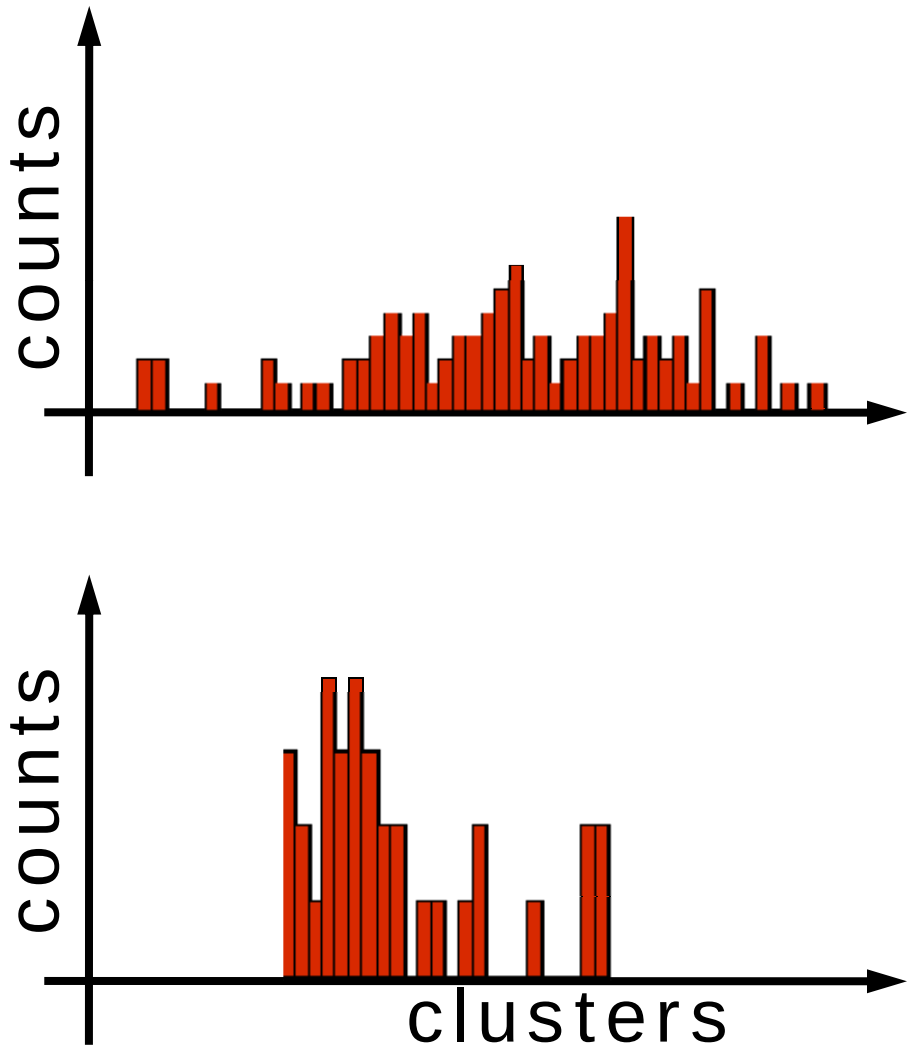
propagating quantitative peptide information to proteins

- each peptide belongs to ONE protein and to ONE cluster!
- the distribution of peptides derived from one particular protein over the various peptide clusters yields a protein signature.

propagating quantitative peptide information to proteins (cont'd)



propagating quantitative peptide information to proteins (cont'd)



Protein-protein distances can be calculated by e.g. ***earth mover's distances***

...and does it work?

Description	PSS
CCNB1: G2/mitotic-specific cyclin-B1	0
TK1: Thymidine kinase cytosolic	2
PRC1: Protein regulator of cytokinesis 1	6
TPX2: Targeting protein for Xklp2	7
NUSAP: Nucleolar/spindle-assoc. protein 1	12
PLK1: Serine/threonine-protein kinase	24
CKAP2: Cytoskeleton-associated protein 2	399
AURKA: Serine/threonine-protein kinase 6	548
CDCA5: Sororin	1565
DNMT1: DNA methyltransferase 1	1598
GTSE1: G2 and S phase-expressed protein 1	1724
Confirmed proteins in top 1% ranks	5/10
Ratio of confirmed proteins ($q = 1\%$)	20.8%
Enrichment factor ($q = 1\%$)	50.9

PSS: Kirchner *et al.* Bioinformatics, 2010, 26, 77pp
PCP: Andersen *et al.* Nature, 2003, 426, 570pp

summarizing the next step: an appropriate statistical framework for the analysis of PSS data

BIOINFORMATICS ORIGINAL PAPER

Vol. 26 no. 1 2010, pages 77–83
doi:10.1093/bioinformatics/btp607

Gene expression

Computational protein profile similarity screening for quantitative mass spectrometry experiments

Marc Kirchner[1,2,†], Bernhard Y. Renard[1,3,†], Ullrich Köthe[3], Darryl J. Pappin[4], Fred A. Hamprecht[1,3], Hanno Steen[1,2,‡,*] and Judith A. J. Steen[5,‡]

[1]Department of Pathology, Proteomics Center, Children's Hospital Boston, [2]Department of Pathology, Harvard Medical School, Boston, MA, USA, [3]Interdisciplinary Center for Scientific Computing, University of Heidelberg, Heidelberg, Germany, [4]Cold Spring Harbor Laboratory, Cold Spring Harbor, NY and [5]Department of Neurobiology, Harvard Medical School and T. M. Kirby Neurobiology Center, Children's Hospital, Boston, MA, USA

Received on May 22, 2009; revised on September 14, 2009; accepted on October 7, 2009

Advance Access publication October 27, 2009

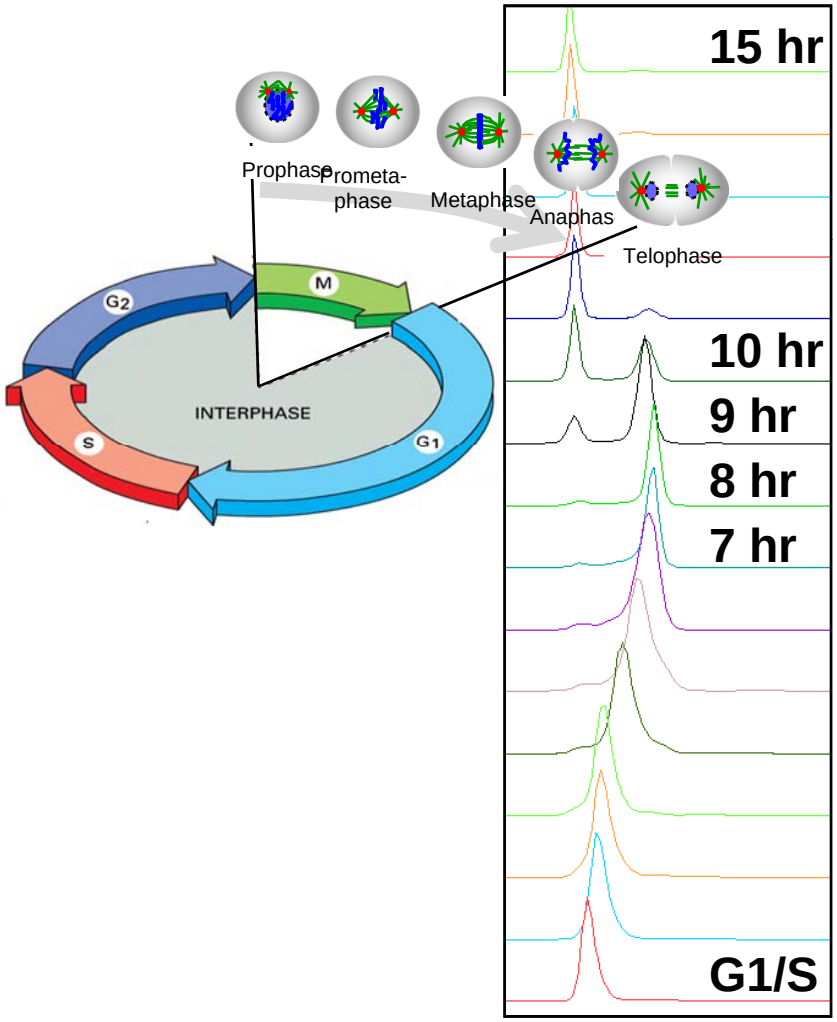
Associate Editor: Burkhard Rost

summarizing the next step: an appropriate statistical framework for the analysis of PSS data

- Normalization leads to the ***loss of one degree of freedom***.
- Transformed/simplicial data are ***not normally distribution***.
- ***Dirichlet statistics*** have to be applied.
- Successful propagation of quantitative peptide information to proteins.
- Working with normalized data and appropriate statistics yielded a ***50-fold enrichment of APC/C substrates*** among the 1 % of proteins best matching the cyclin B profile.
- Moving forward: moving beyond prototyping…

APC/C substrate screen 2009/2010: all new and improved!

- New synchronization protocol to minimize dead cells.
- Focus on mitosis.
- Higher temporal resolution.
- Increased multiplexing by using 6-plex TMT instead of original 4-plex iTRAQ.
- LC/MS using an LTQ-Orbitrap instead of MALDI TOF/TOF.
- OFFGel fraction.
- 3 biological and 3 technical replicates.

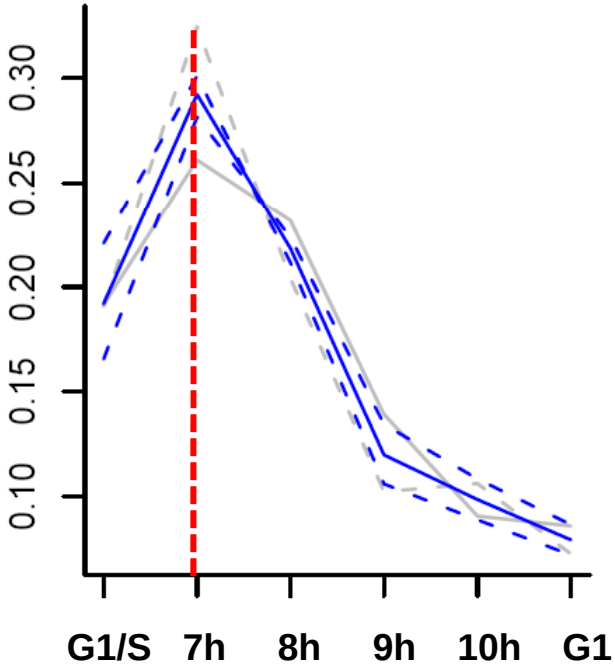


2nd generation APC/C substrate screen: in numbers

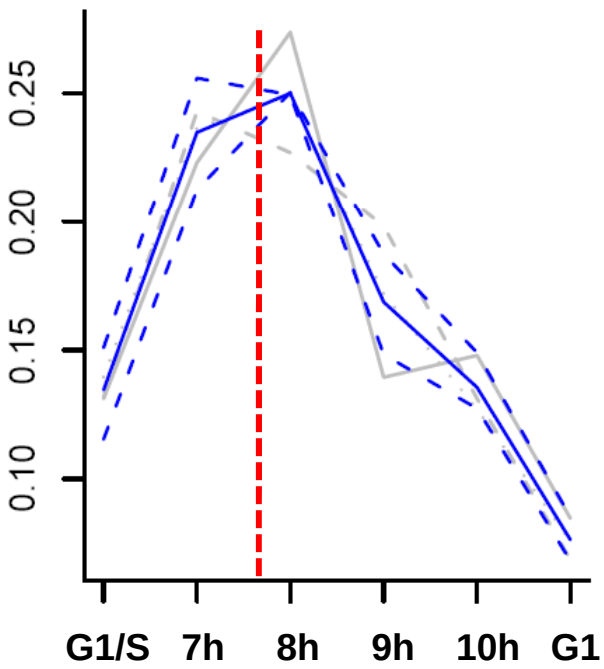
- ✓ ~200.000 peptides w/ complete reporter ion traces (x 10)
- ✓ ~6000 proteins identified and quantified (x 2.5)
- ✓ ~50 % of all known APC/C substrates identified (x 2)

a closer look at the new dataset

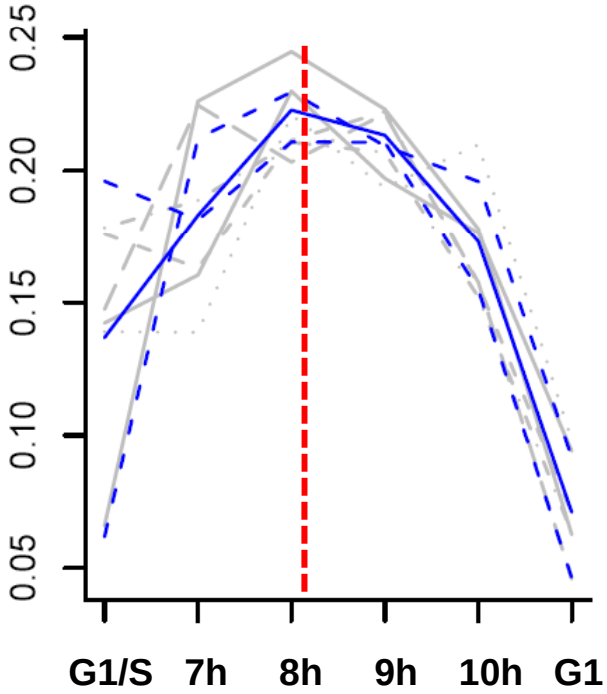
Only Cdc20 substrates



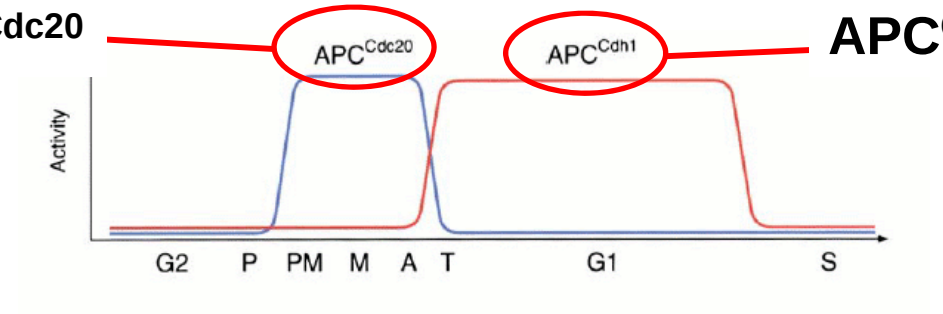
Mixed Cdc20 and Cdh1 substrates



Only Cdh1 substrates



APC^{Cdc20}



APC^{Cdh1}

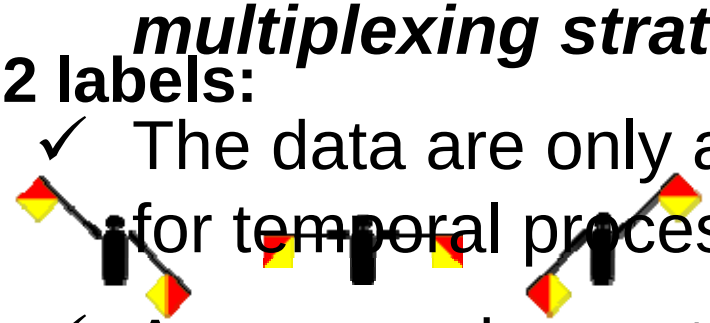
summary

- ✓ ***Quantitative proteomics*** is capable of identifying APC/C substrates.
- ✓ ***Protein Profile Similarity Screening*** is very useful for and successful in identifying co-regulated proteins.
- ✓ PPSS provides the ***appropriate statistical framework*** to account for normalization-related problems of relative quantitative information.
- ✓ Although shown for temporal processes, this concept is also applicable to e.g. changing concentration of various inhibitors, etc…
- ✓ Carefully planned quantitative proteomics experiments in combination with appropriate bioinformatics tools enables true '***functional proteomics***'.

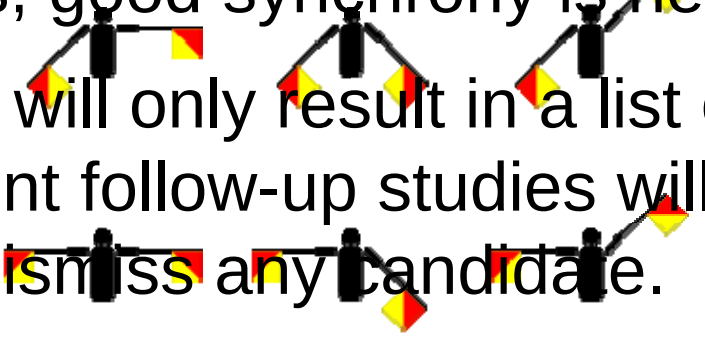
some additional considerations…

- ✓ To use PPSS at its full potential, data from ***multiplexed*** quantitative proteomics experiments are necessary.
- ✓ n labels result in 3^{n-1} ***fundamental profiles:***
- ✓ ***We need more non-TMT and non-iTRAQ-based multiplexing strategies!***

2 labels:



3 labels:



etc…

- ✓ The data are only as good as your biochemical input; for temporal processes, good synchrony is necessary
- ✓ Any screening method will only result in a list of candidates; independent follow-up studies will be necessary to confirm/dismiss any candidate.

acknowledgements

The Prototype:

Phil Ross (AB)

Darryl Pappin (AB)

Ken Parker (AB)

Flavio Monigatti (CHB)

Marc Kirchner (U. Heidelberg)

Sven Liffers (CHB)

Amit Tzur (HMS)

Marc Kirschner (HMS)

Judith Steen (CHB)

The Dirichlet Statistics:

Marc Kirchner (CHB)

Bernhard Renard (U. Heidelberg)

Ullrich Koethe (U. Heidelberg)

Fred Hamprecht (U. Heidelberg)

Judith Steen (CHB)

The 2009/2010 Model:

Dominic Winter (CHB)

Marc Kirchner (CHB)

Sasha Singh (CHB)

Nurhan Ozlu (CHB)

Judith Steen (CHB)