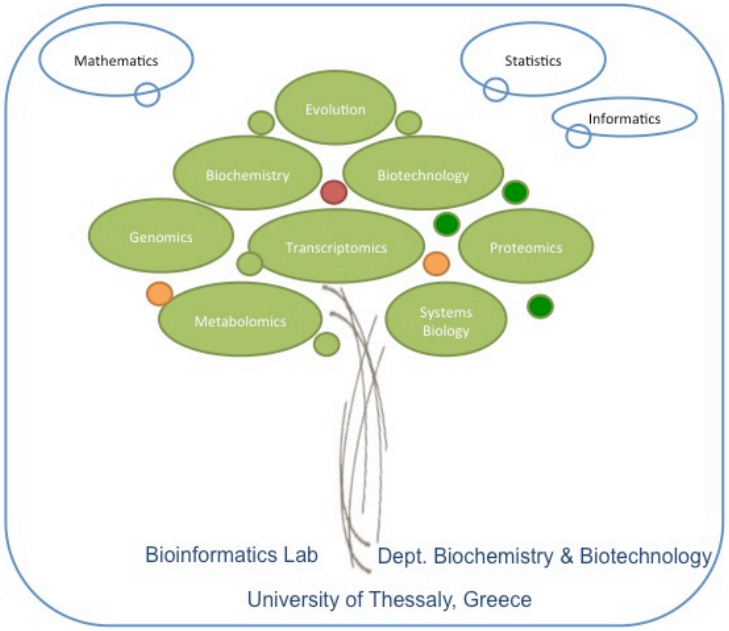
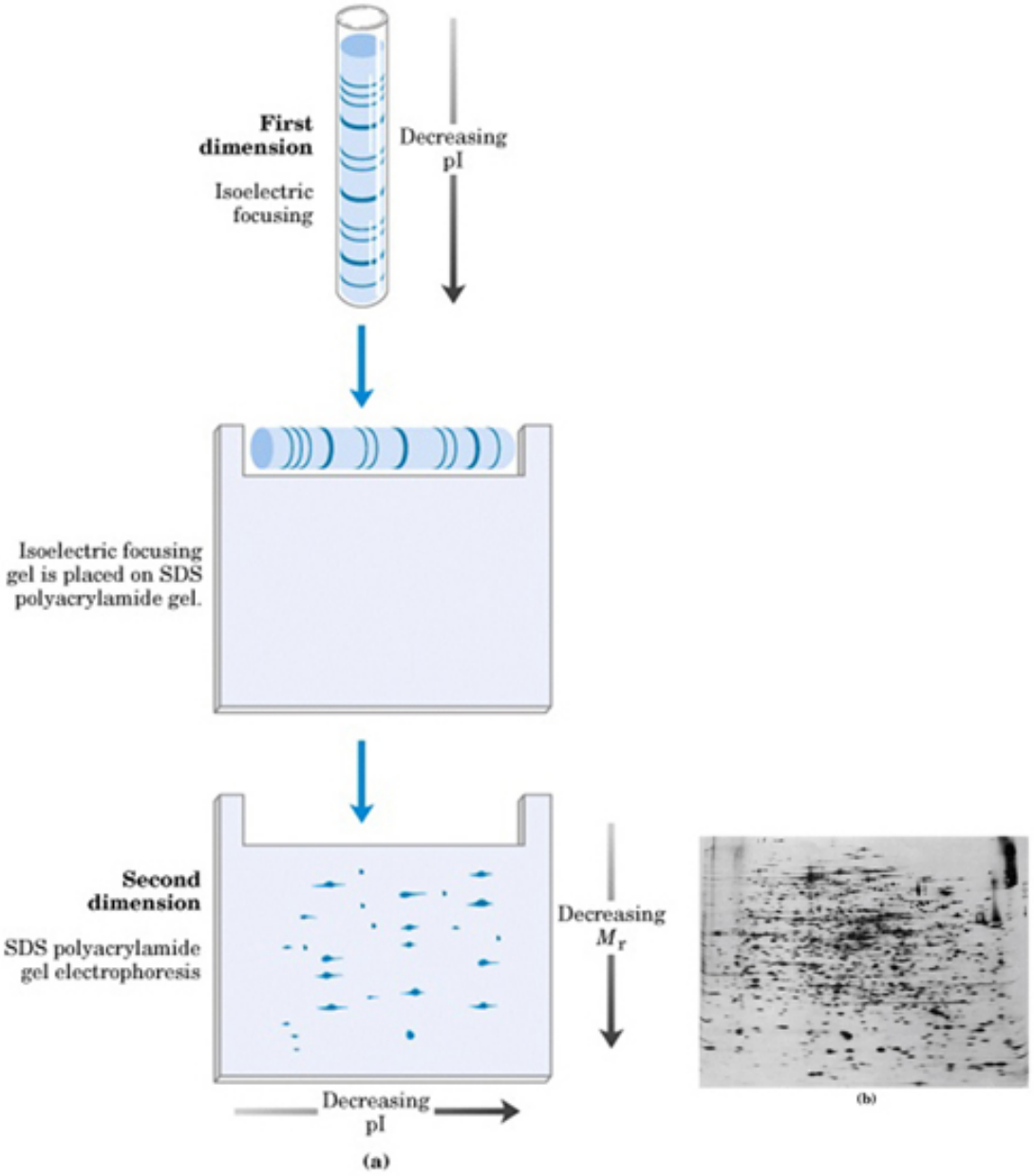
Βασικές αρχές Πρωτεομικής και Φωσφοπρωτεομικής

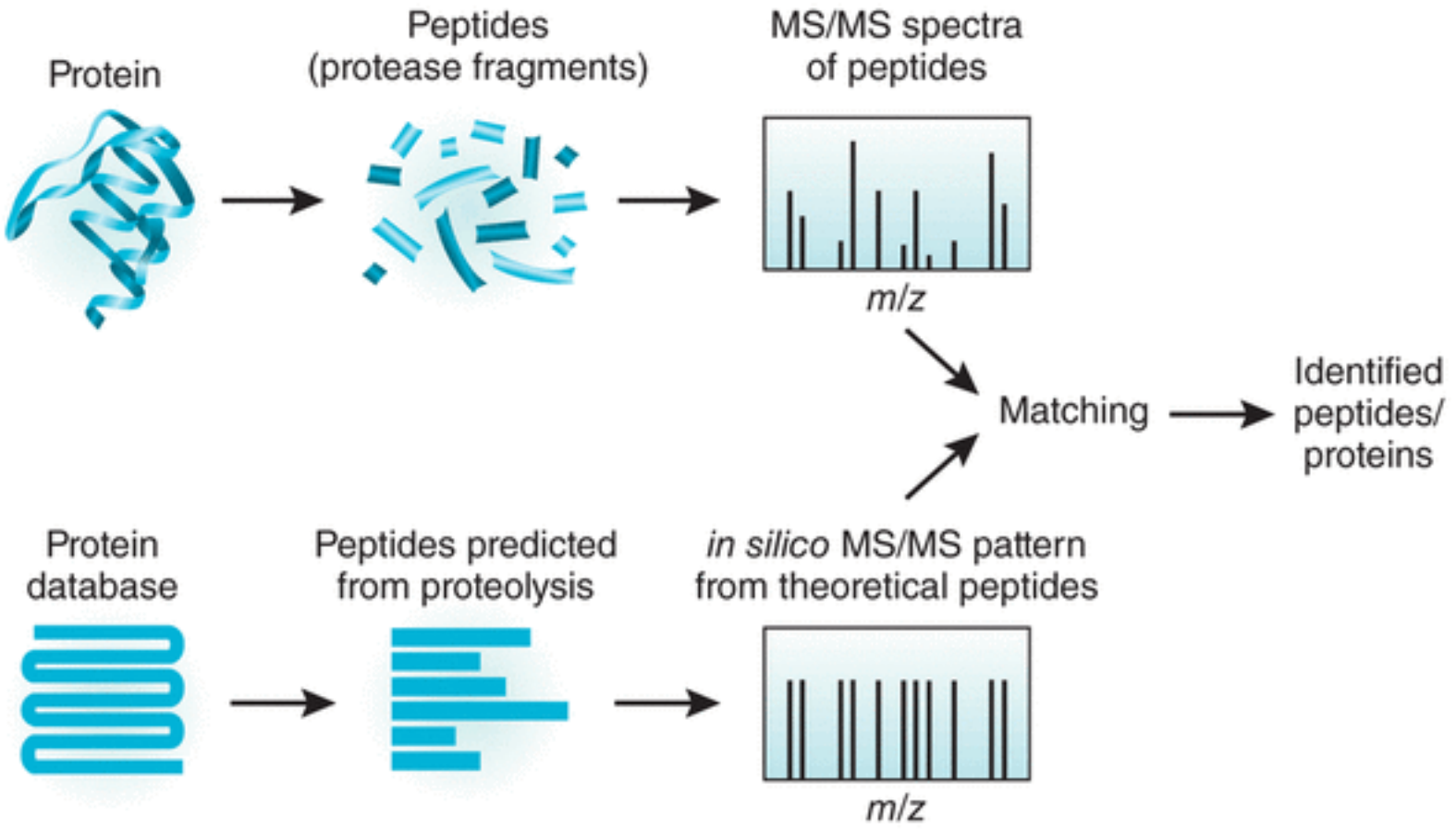
Γρηγόρης Αμούτζιας
Αν. Καθηγητής Βιοπληροφορικής με έμφαση στη Μικροβιολογία,
Τμήμα Βιοχημείας και Βιοτεχνολογίας,
Πανεπιστήμιο Θεσσαλίας



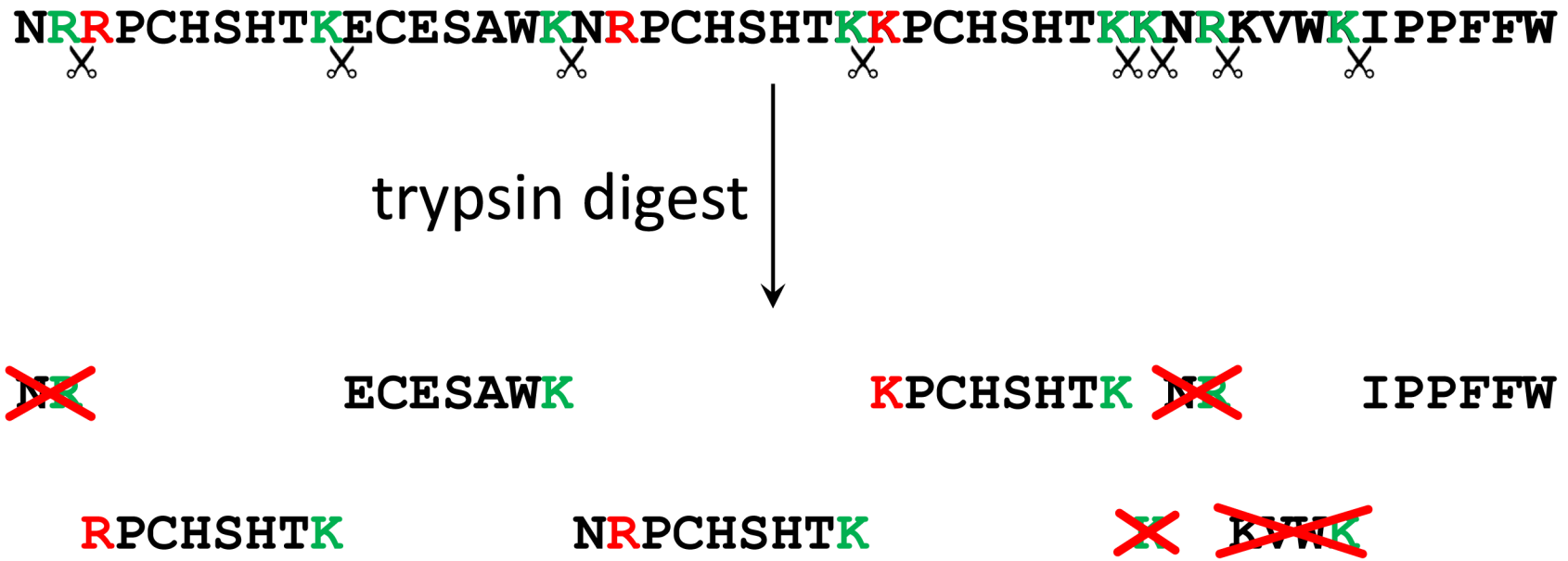
Two-D Gels


First dimension
Isoelectric focusing
Decreasing pI
Isoelectric focusing gel is placed on SDS polyacrylamide gel.
Second dimension
SDS polyacrylamide gel electrophoresis
Decreasing M_r
Decreasing pI
(a)
(b)

Φασματομετρία μάζας για πρωτεομική

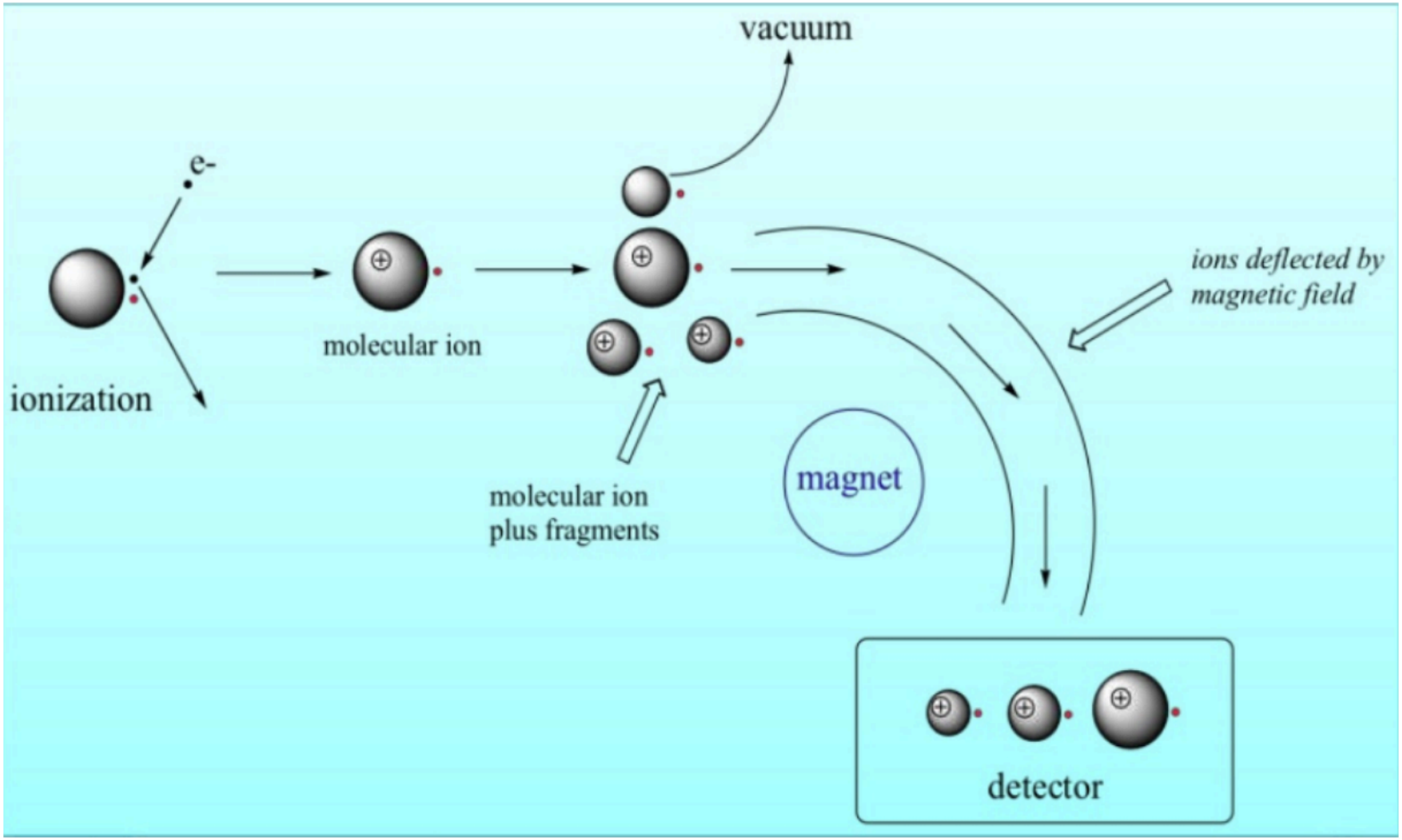


Πέψη με τρυψίνη

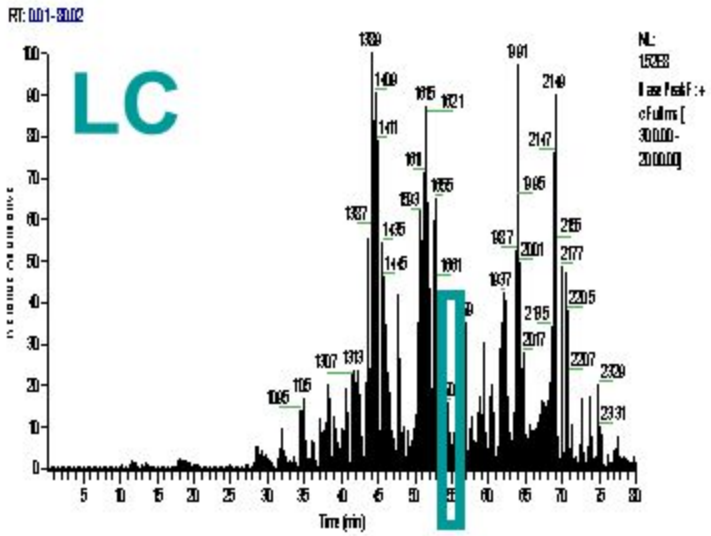


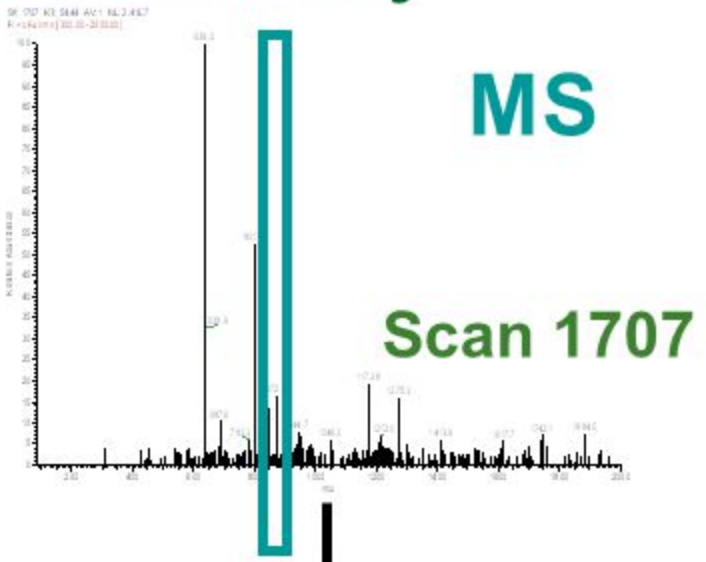
- Κάποιες φορές η τρυψίνη μπορεί να μην κόβει κάποιες θέσεις, οπότε δημιουργούνται αλληλεπικαλυπτόμενα κομμάτια.
- Επίσης, μπορεί να χρησιμοποιηθούν περισσότερες από μία διαφορετικές πρωτεάσες για να έχουμε καλύτερη κάλυψη της πρωτεινης

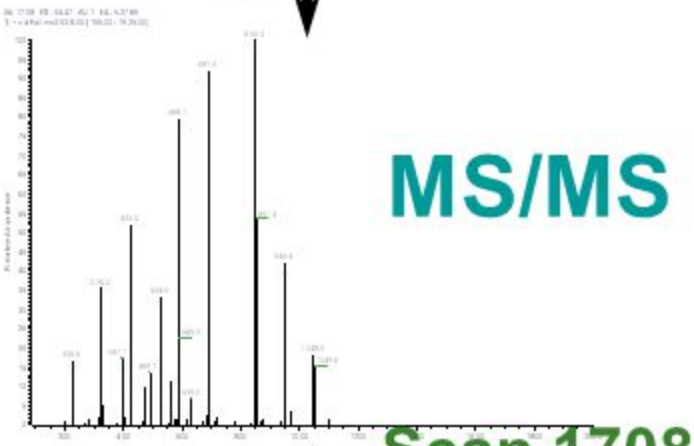
Φασματομετρία μάζας – βασικές αρχές

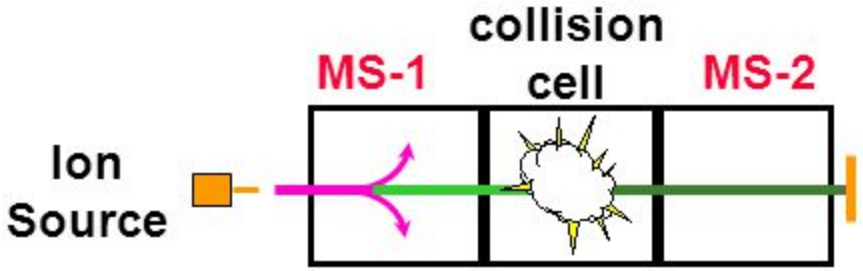


Tandem Mass Spectrometry

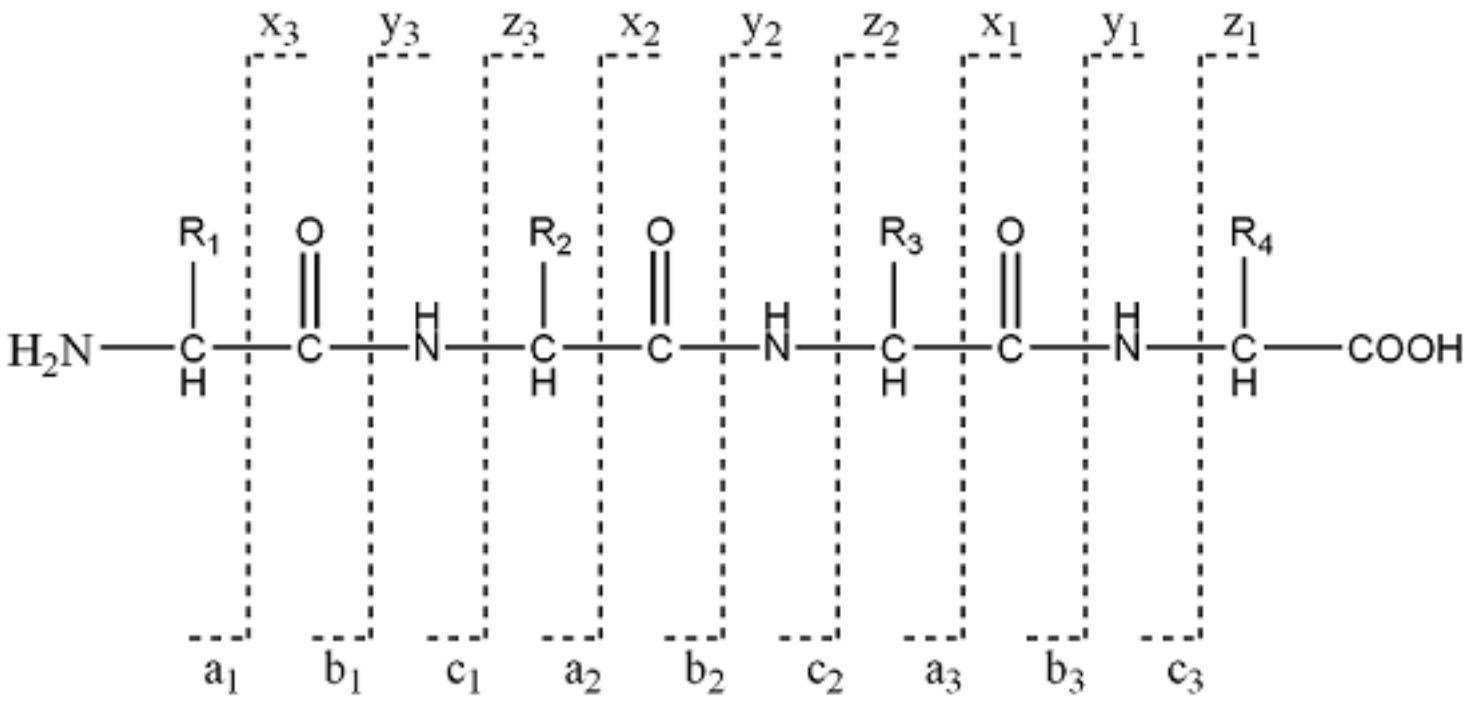








Θραύση πεπτιδίου σε διάφορα ιόντα

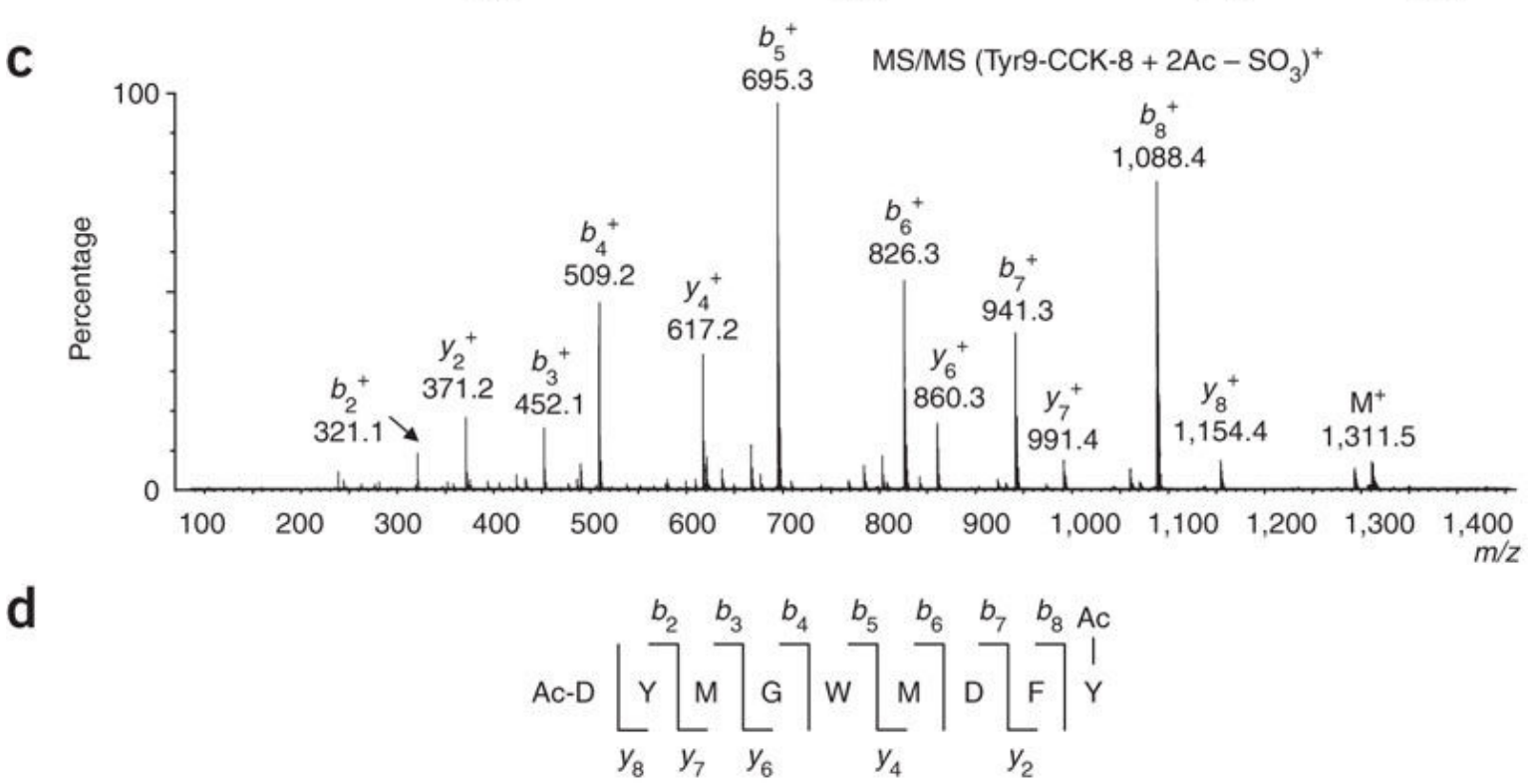


Θραύση πεπτιδίου σε διάφορα ιόντα

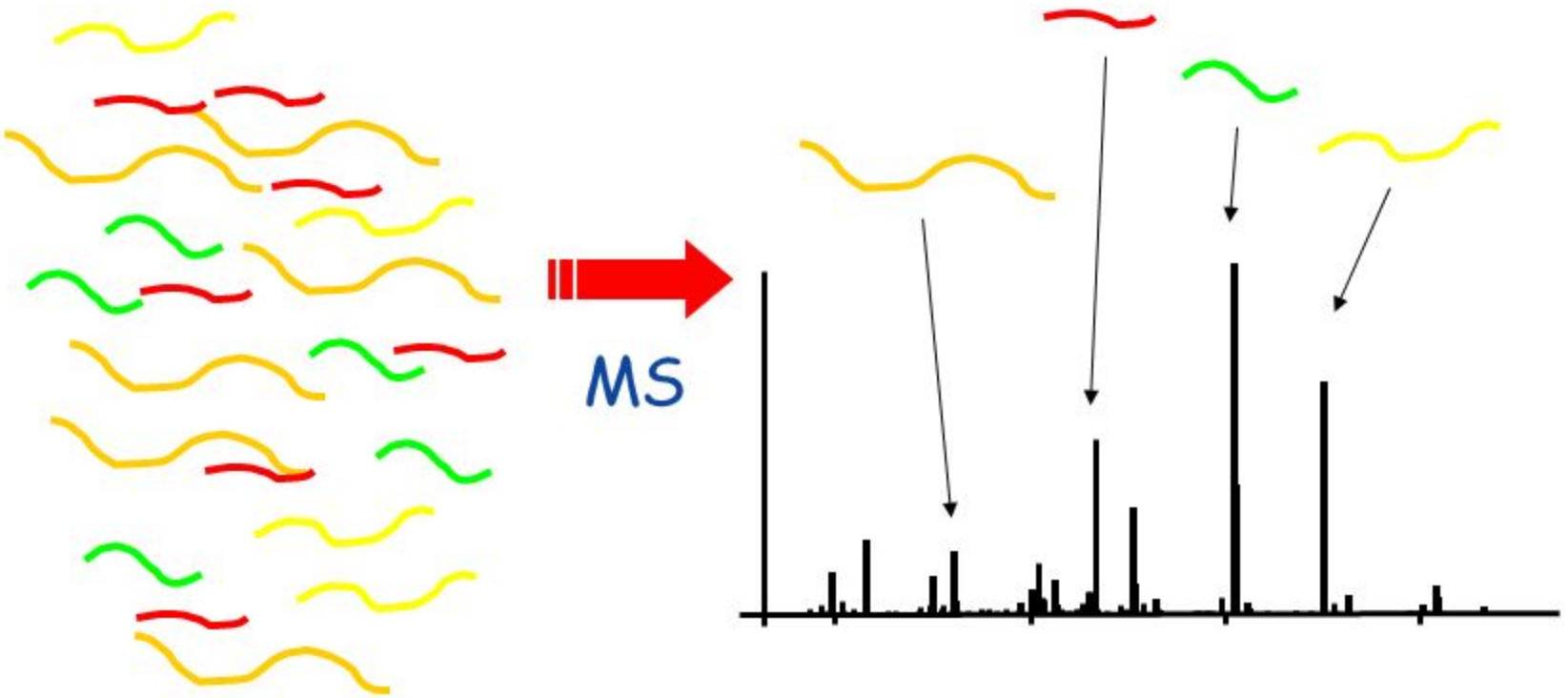
Peptide: S-G-F-L-E-E-D-E-L-K

MW	ion			ion	MW
88	b_1	S	GFLEEDELK	y_9	1080
145	b_2	SG	FLEEDELK	y_8	1022
292	b_3	SGF	LEEDELK	y_7	875
405	b_4	SGFL	EEDELK	y_6	762
534	b_5	SGFLE	EDELK	y_5	633
663	b_6	SGFLEE	DELK	y_4	504
778	b_7	SGFLEED	ELK	y_3	389
907	b_8	SGFLEEDE	LK	y_2	260
1020	b_9	SGFLEEDEL	K	y_1	147

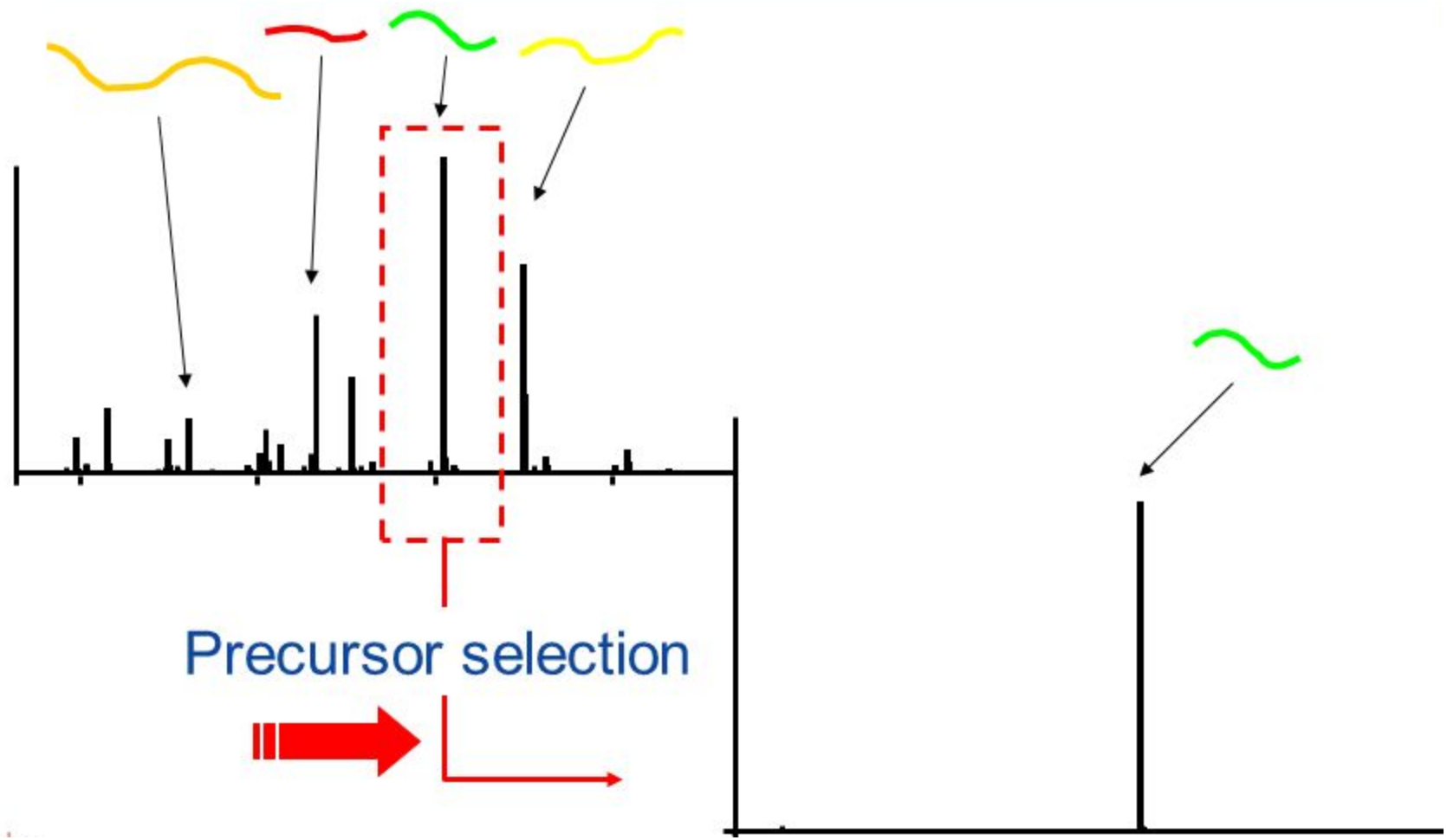
Φάσμα ιόντων



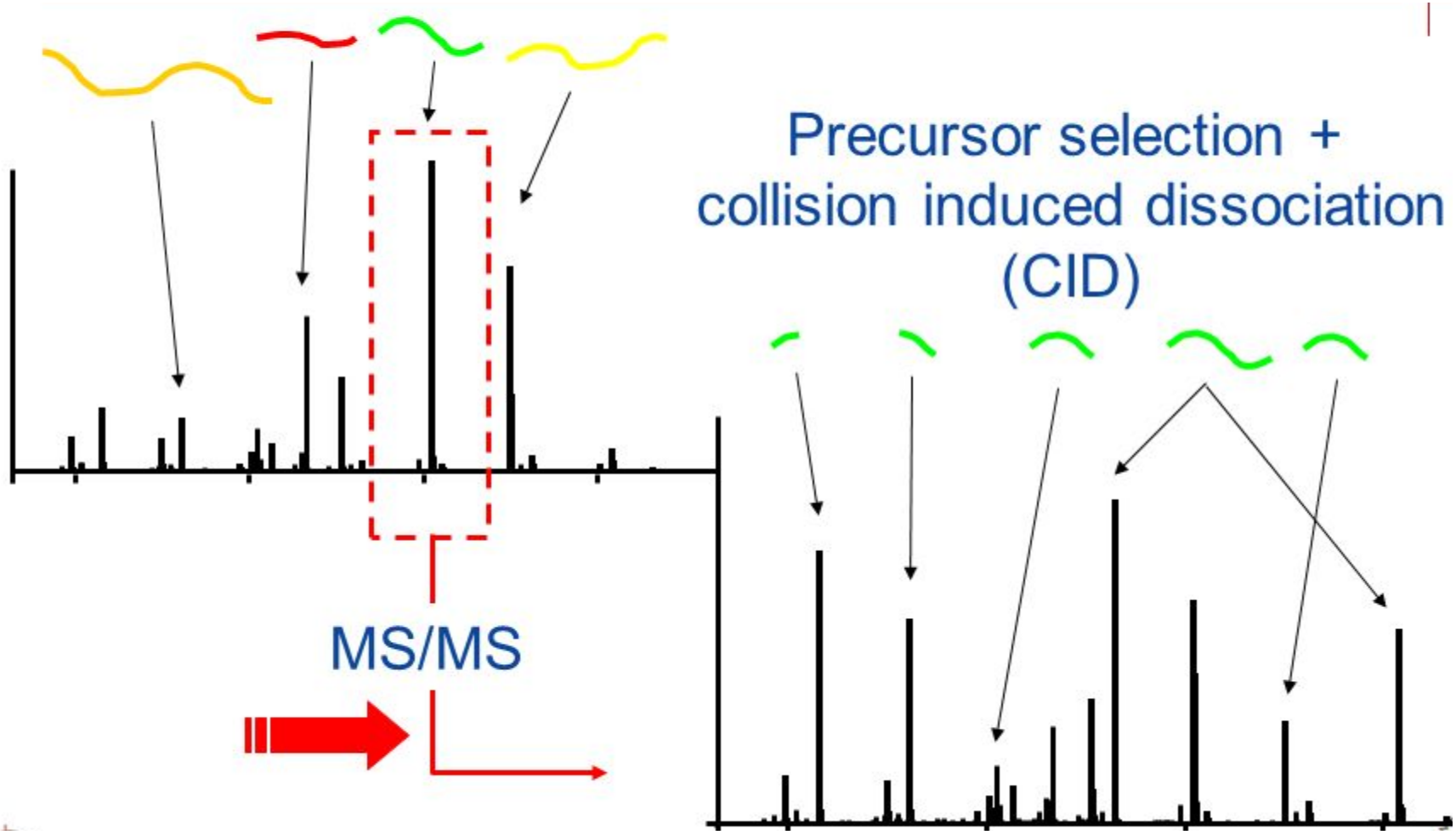
Φασματομετρία μάζας – MS1



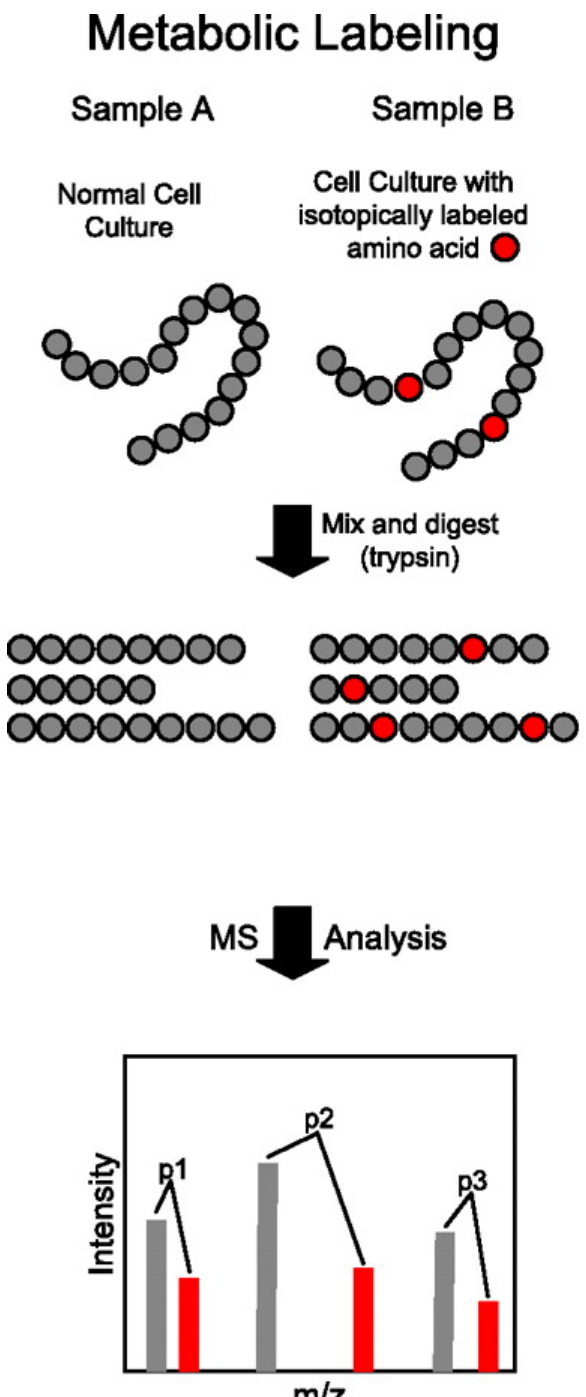
Φασματομετρία μάζας – MS1



Φασματομετρία μάζας – MS2



Ισότοπα αμινοξέων για ποσοτικοποίηση



Proteomics software

- MASCOT
- MAXQUANT

- Χρειαζόμαστε ένα score/πιθανότητα για τον σωστό εντοπισμό του πεπτιδίου
- Χρειαζόμαστε ένα score/πιθανότητα για τον σωστό εντοπισμό της πρωτεινης
- Χρειαζόμαστε ένα score/πιθανότητα για τον σωστό εντοπισμό μιας μετα-μεταφραστικής τροποποίησης, όπως πχ. φωσφορυλίωση

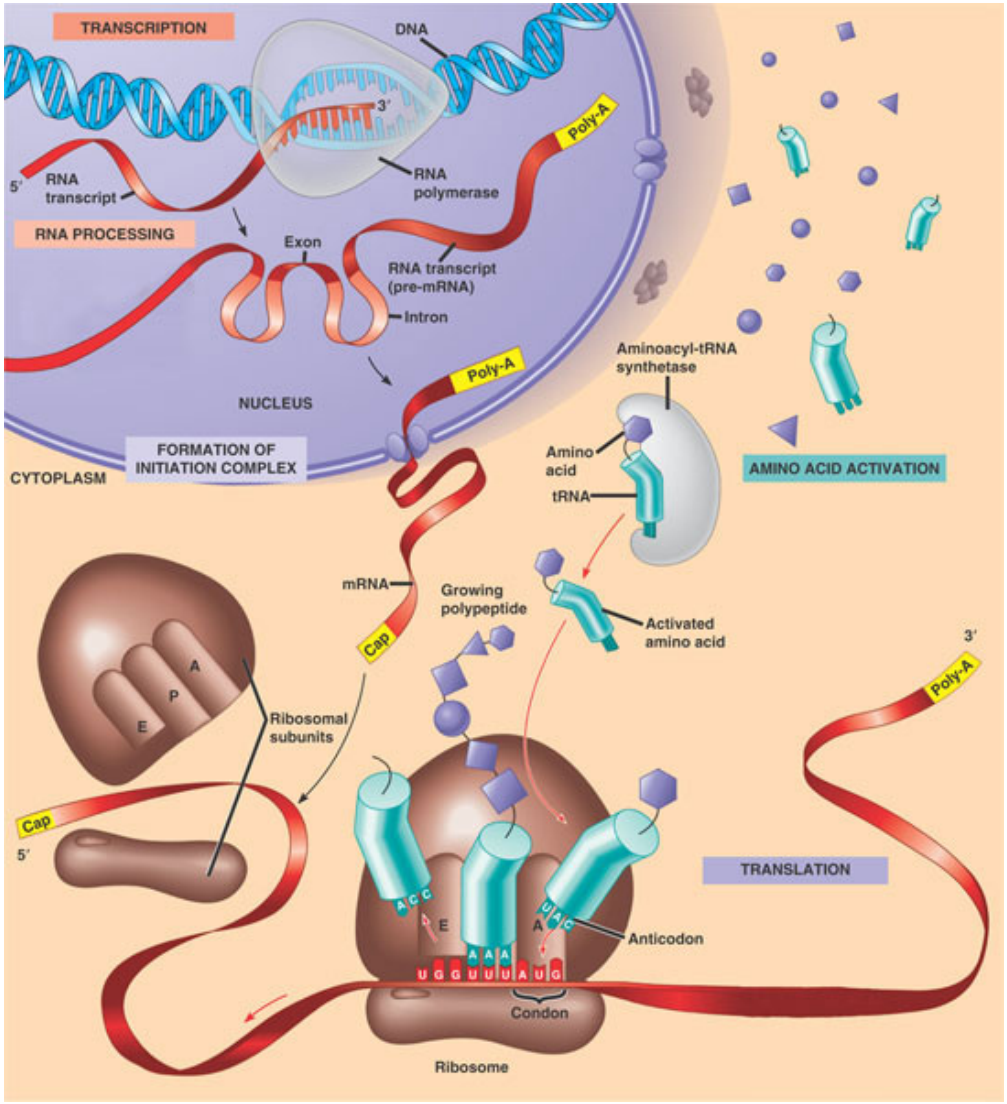
The eukaryotic phosphoproteome through the bioinformatics prism:

evaluation and properties

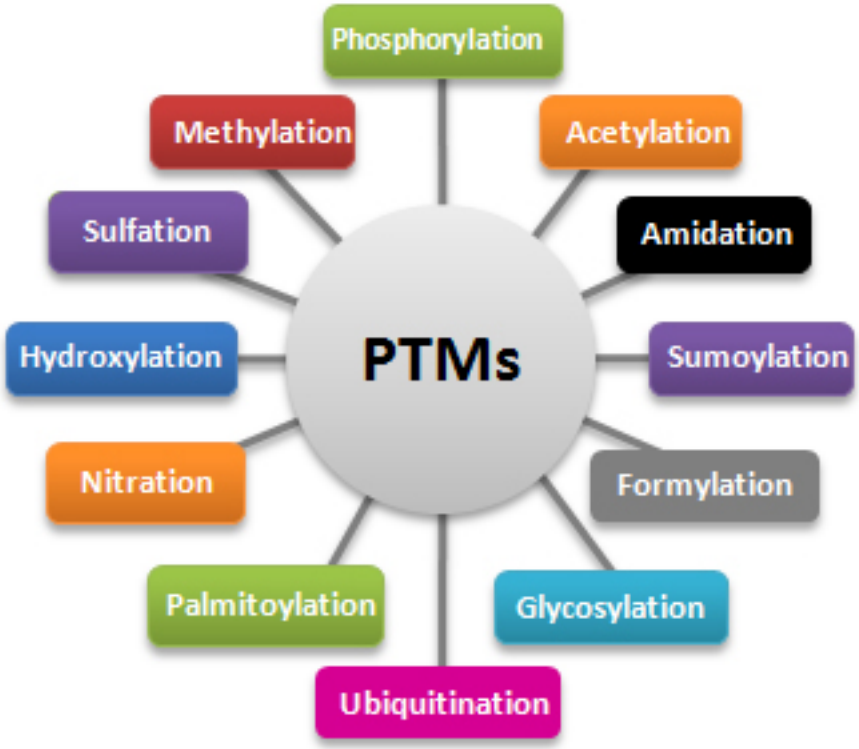
Grigoris Amoutzias

Assistant Professor of Bioinformatics in Genomics

Many levels of gene regulation

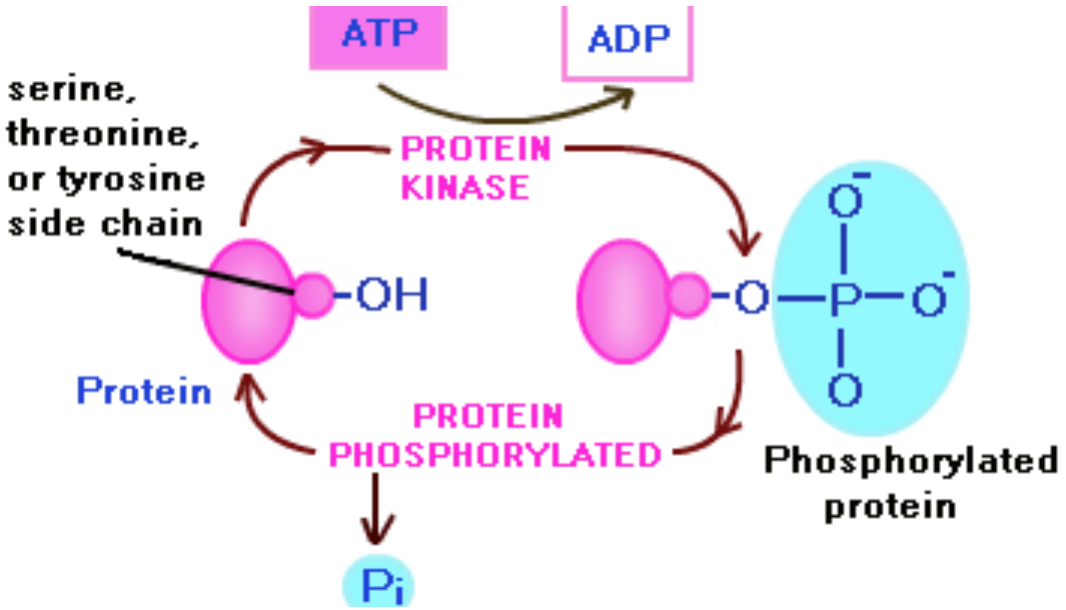


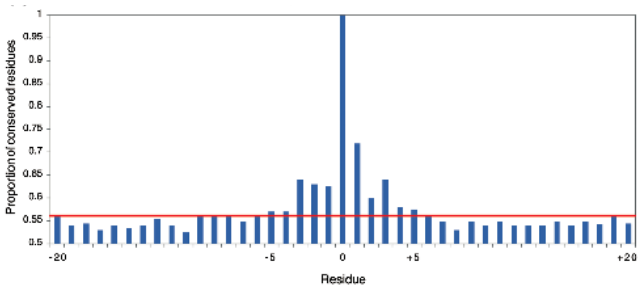
Post-translational regulation: fast and energy efficient



What is protein phosphorylation

- Addition of a phosphate group on a Serine, Threonine, or Tyrosine, by kinases.
- Amino acid motifs for phosphorylation are short.
- Phosphorylation motifs are known to occur within unstructured and rapidly evolving regions (loops).





Gnad et al., 2007; Genome Biology

What is protein phosphorylation

Acts as a switch

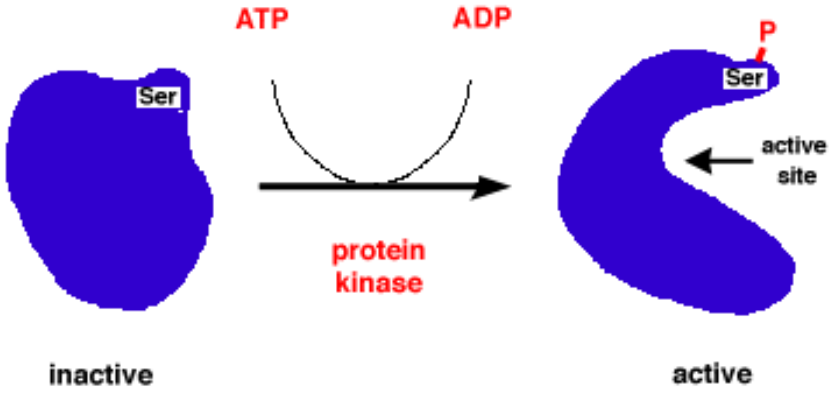
Acts as a dimmer

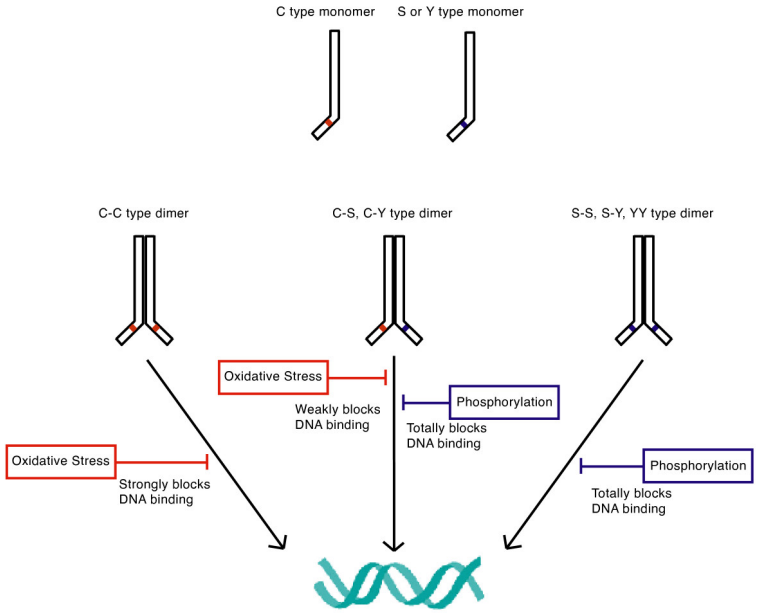
Importance of phosphorylation

Manipulation of molecular pathways and phenotypes, by modifying a small number of phosphorylation sites, via a few point mutations.

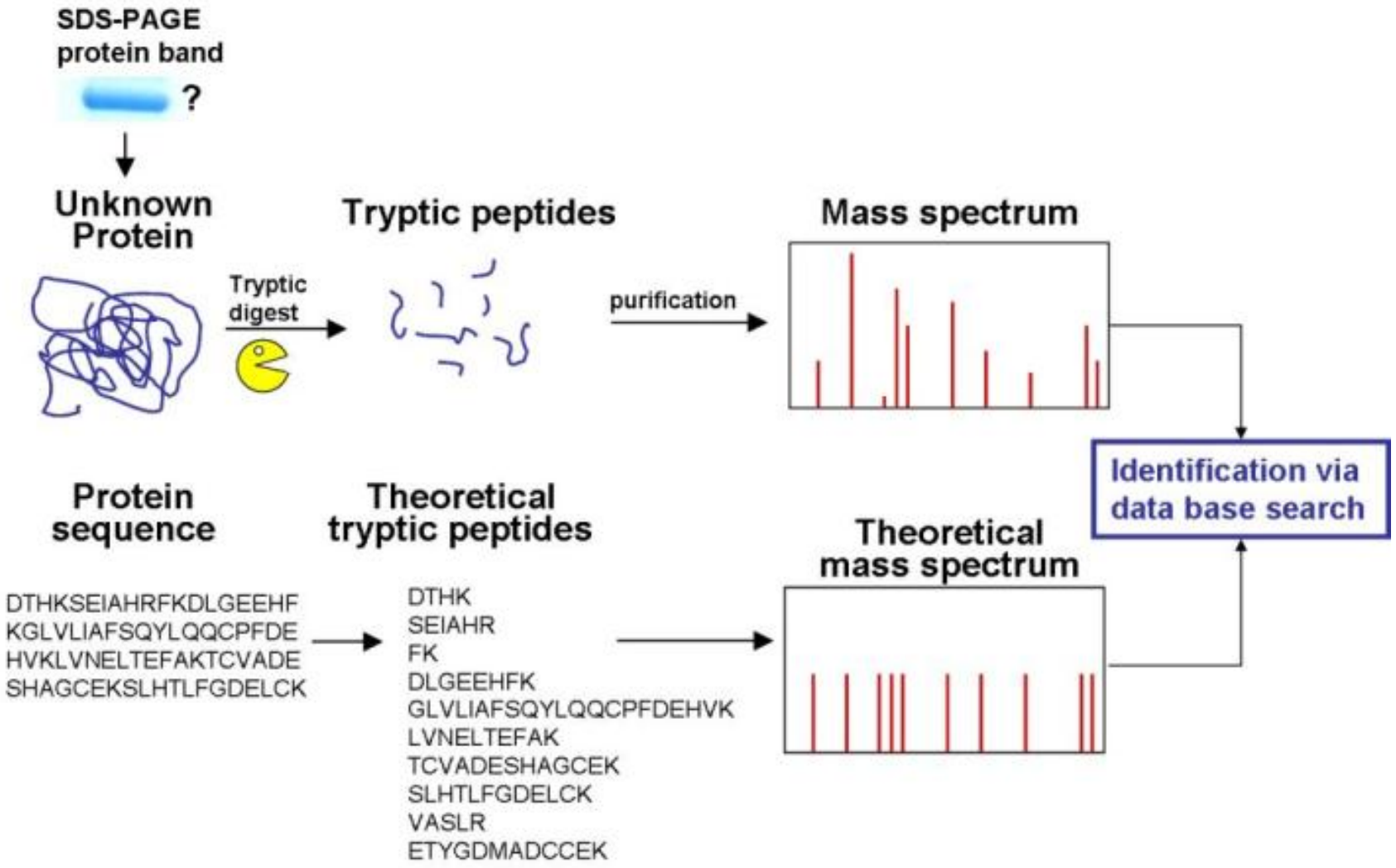
Phosphorylation involved in many diseases.

A point mutation in cdc28 (S42->A) results in decrease of cell size,

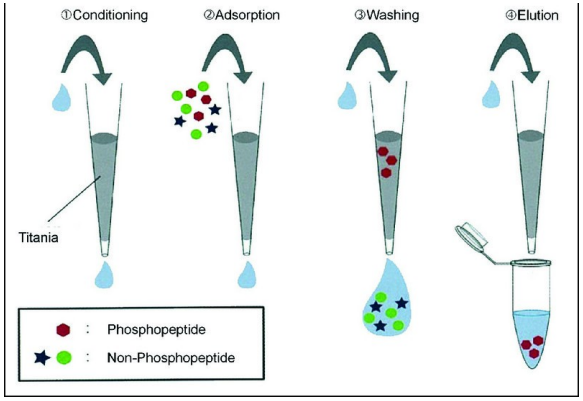




From Low to High-throughput: Enter Proteomics



The era of phosphoproteomics



Phosphopeptide enrichment techniques (IMAC, TiO2)

Very sensitive Mass spectrometers

Bioinformatics

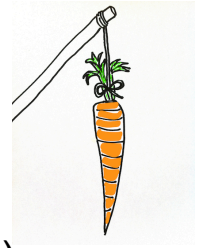
Identifed Sequence	PeptideProphet	Tryptic Ends	Peptide Mass	DeltaCN	# Obs	# Mappings	Links
R.NT*TPYQNNVYNDAIR.D	0.99	2	1862.82	0.21	34	1/1	
R.NTT*PYQNNVYNDAIR.D	0.97	2	1862.82	0.05	1	1/1	
K.GIRPSPLENS*LHR.A	0.99	2	1555.78	0.28	1	1/1	

⊖ Protein/Peptide Sequence

YER118C | SHO1

MSISSKIRPTPRKPSRMATDHSFKMKKFYADPFAISSISLAIVSWVIAIGGSISSASTNE
SFPRFTWWGIVYQFLIICSLMLFYCFDLVDHYRIFITTSIAVAFVYNTNSATNLVYADGP
KKAAASAGVILLSIINLIWILYYGGDNASPTNRWIDSFSIKGIRPSPLENSLHRARRRGN
RNTTPYQNNVYNDAIRDSGYATQFDGYPQQQPSHTNYVSSTALAGFENTQPNTSEAVNLH
LNTLQQRINSASNAKETNDNSNNQTNTNIGNTFDTDFSNGNTETTMGDTLGLYSDIGDDN
FIYKAKALYPYDADDDDAYEISFEQNEILQVSDIEGRWWKARRANGETGIIPSNYVQLID
GPEEMHR

Motivation



- Many high-throughput phosphoproteomic datasets (with various technologies) have come out, but no thorough comparative evaluation yet.

- Previous studies: each technology has its biases.
  - Capture different (but also overlapping) sub-space of the entire phosphoproteome.

- Questions arising, related to the high sensitivity of the MS-technology.
  - Low stoichiometry phosphorylations (Lienhard)
  - Non-functional psites (Landry)
  - Correct detection/localization of p-sites
  - Same dataset, different software: ~30% overlap in results

- How good are the current phosphoproteomic technologies?
- Are conclusions of previous studies, robust, or strongly affected by biases?
- How can we filter the data and obtain a reliable phosphoproteome?
- What are the general properties of a model (the yeast) phosphoproteome?

⌘ *Author's Choice*

Evaluation and Properties of the Budding Yeast Phosphoproteome*[S]

Grigoris D. Amoutzias‡§¶, Ying He¶‖, Kathryn S. Lilley‡, Yves Van de Peer¶‖, and Stephen G. Oliver‡**

Yeast as a model organism

- Large number of phosphoproteomics experiments under a reasonably wide range of conditions.
- Unicellular organism.
- Large fraction (80%) of the predicted yeast proteome expressed and detected (by MS based methods) under normal laboratory growth conditions.
- A wealth of relevant functional genomic information available for the organism, including data on
  - protein abundance
  - half-lives,
  - number of kinases targeting a given protein.
- Many essential yeast genes may be complemented by human orthologs.
- A model for pathogenic fungi.
- All of these factors should assist in an in-depth bioinformatics analysis of the yeast phosphoproteome.

Part A:
Quality of the datasets

Contribution of each dataset

- 12 HTP phosphoproteomic datasets
- 99% correct phosphopeptide identification
- 99% correct psite localization
- 9783 p-sites found in 2374 phosphoproteins

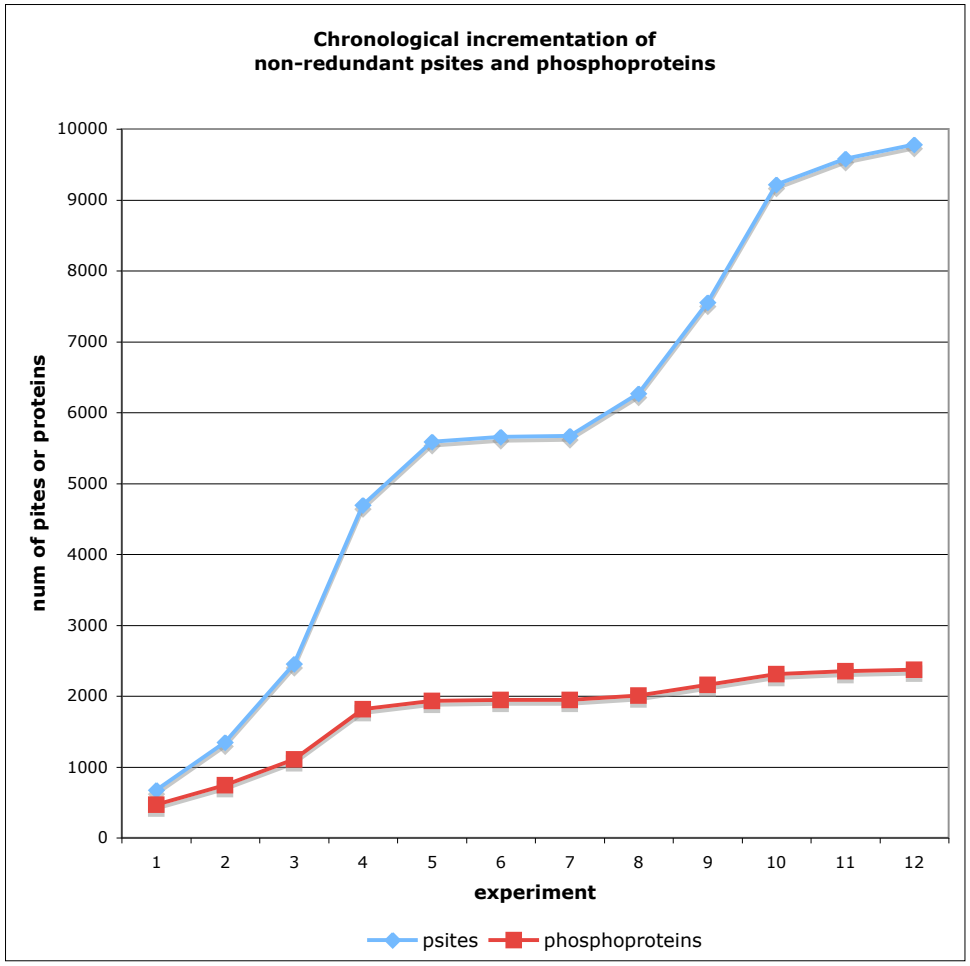
- If a single dataset dominates the compendium, its biases will affect our general conclusions.

- **No single dataset dominates the compendium.**
  - Removal of each dataset resulted in
    - 0-16% reduction of non-redundant p-sites
    - 0-11% reduction of non-redundant phosphoproteins

Overlap among experiments

- The 12 datasets overlap with each other in a statistically significant manner (chi-squared $p < 0.05$).
- For any two experiments
  - ~12% of p-sites are shared.
  - ~28% of phosphoproteins are shared.
- If not, it would be a reason for concern, some datasets would be of questionable quality and would need to be removed.
- Two experiments from different groups, but on similar biological conditions (alpha-factor treated cells), had a much lower overlap (11% of p-sites & 31% of phosphoproteins) between them than two experiments of the same group that were performed in two different phases of the cell cycle (28% & 54% respectively)
- **Protocol is very important**

Saturation of the compendium



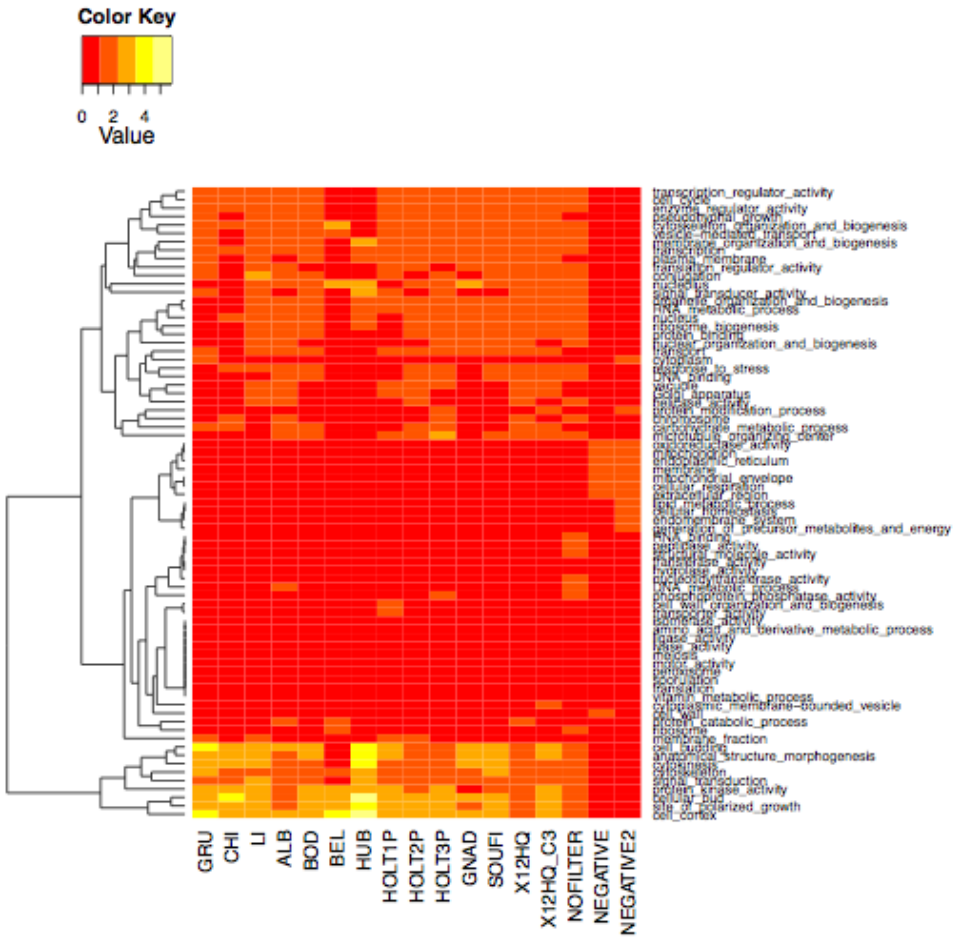
This compendium (12HQ) found:

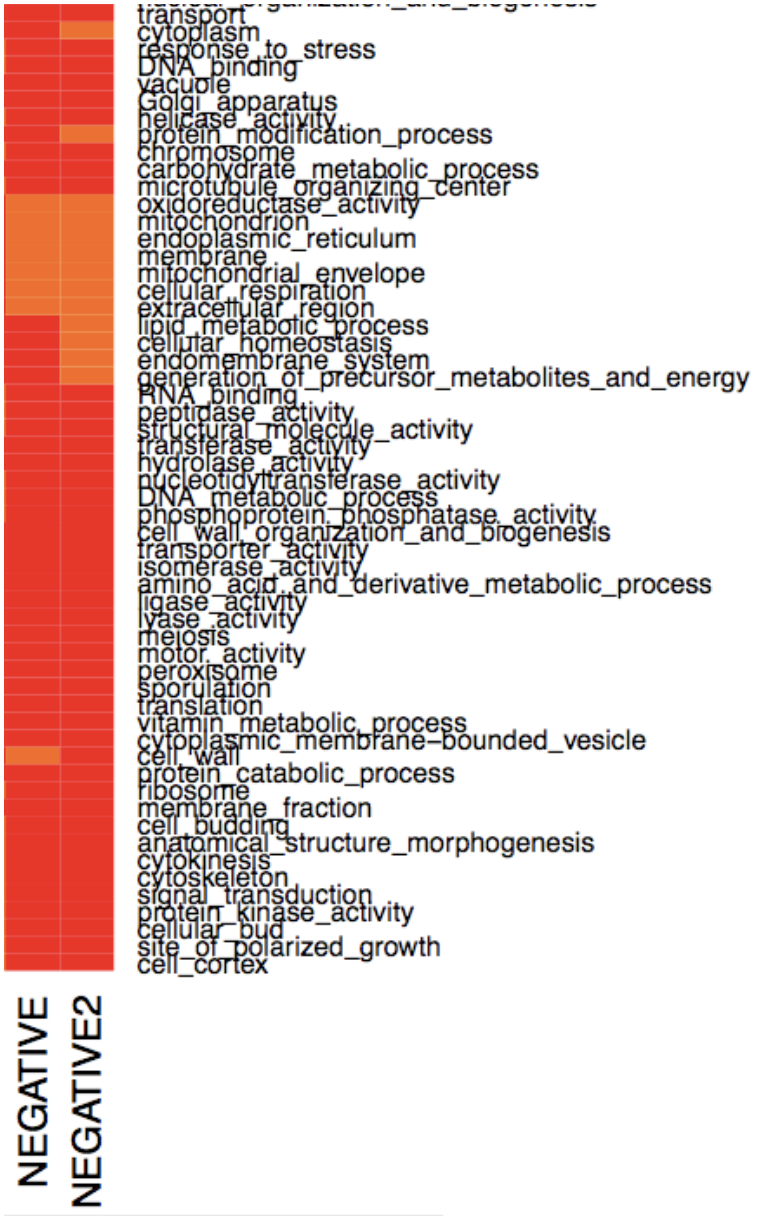
- 27% (131/480) of the PhosphoGrid p-sites
- 85% (122/144) of the PhosphoGrid phosphoproteins

The non-phosphoproteome

- No evidence for phosphorylation in any of the 12 HTP experiments (even with no filtering applied)
- 2219 ORFs.

The non-phosphoproteome GO-Slim analysis





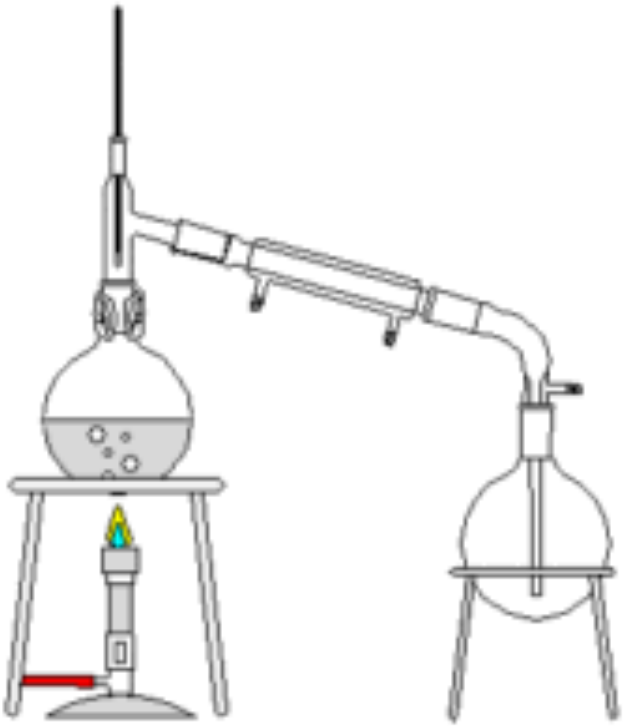
- Mitochondria - Verified by mitochondria-targeted phosphoproteomics (bacterial origin)
- Membranes
- Cell-wall
- Endoplasmic reticulum
- Extracellular space

Could it be an artefact?

- The non-phosphoproteome could be an artefact because of:
  - Inherent undetectability by MS-proteomics
  - Peptide coverage
  - Protein abundance
  - Protein half-life
  - Different properties (length or relative-charge) of digested peptides

The non-phosphoproteome does not appear to be a technical artefact

Part B:
Filtering out “noisy” p-sites



Filtering out “noisy” p-sites

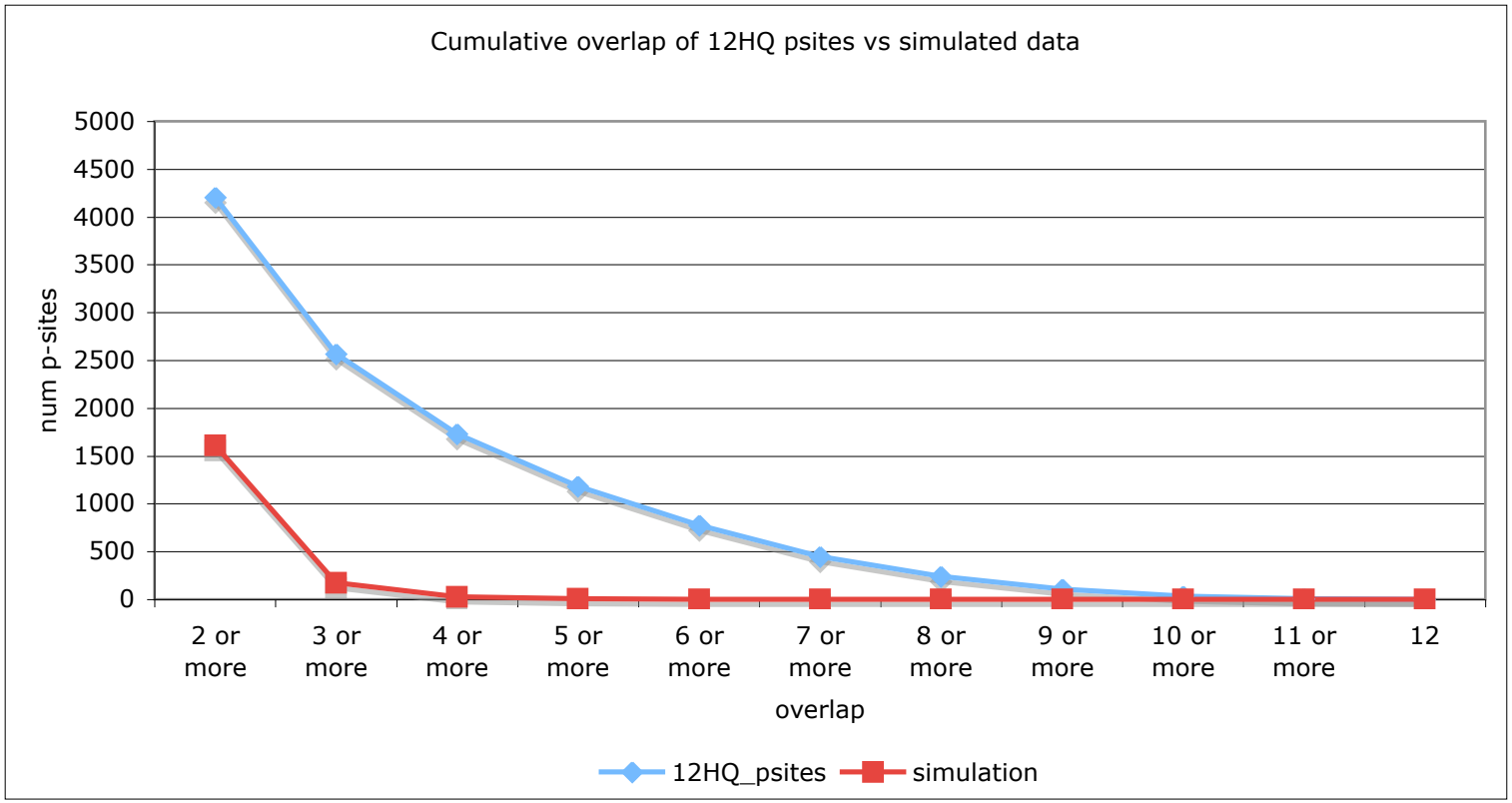
- MS-technologies are very sensitive.
- They could possibly detect low stoichiometry off-target phosphorylations on degenerate motifs (Lienhard, 2008).

- Landry et al., 2009 used evolutionary analyses on smaller HTP-datasets and estimated that up to 65% of p-sites could be non-functional.
- The presence of many experiments allows to address this very important issue.

- We assume that a p-site found in many experiments is more probable to be functional, than “noisy”.
- Five analyses strengthen the validity of the above assumption.

In how many experiments?

- In how many experiments should a p-site have been discovered in order to confidently designate it as functional?
- We simulated the datasets, assuming that all p-sites were assigned in a totally random manner.
- A cutoff of ≥3 seemed stringent.
- We generated a more stringent dataset (12HQ_3x) with 2566 p-sites in 1112 phosphoproteins.



Why so many p-sites found in so many experiments?

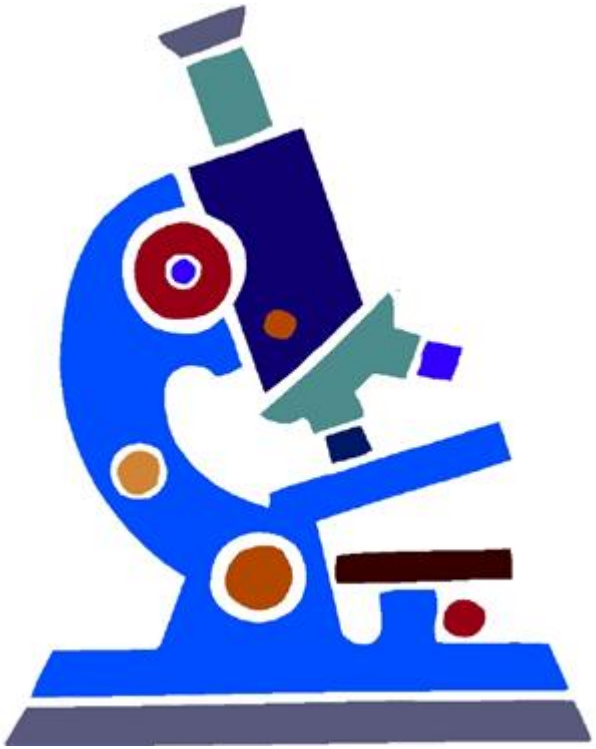
- According to Soufi *et al*. (2009) this could be explained by the asynchronous state of the cell populations in most of the experiments.

- Relatively high overlap (28% and 54% for p-sites and phosphoproteins respectively) in the 2 Holt *et al*. experiments [10], which characterised the phosphoproteome at two different stages of the cell cycle, indicates that this cannot be a complete explanation.

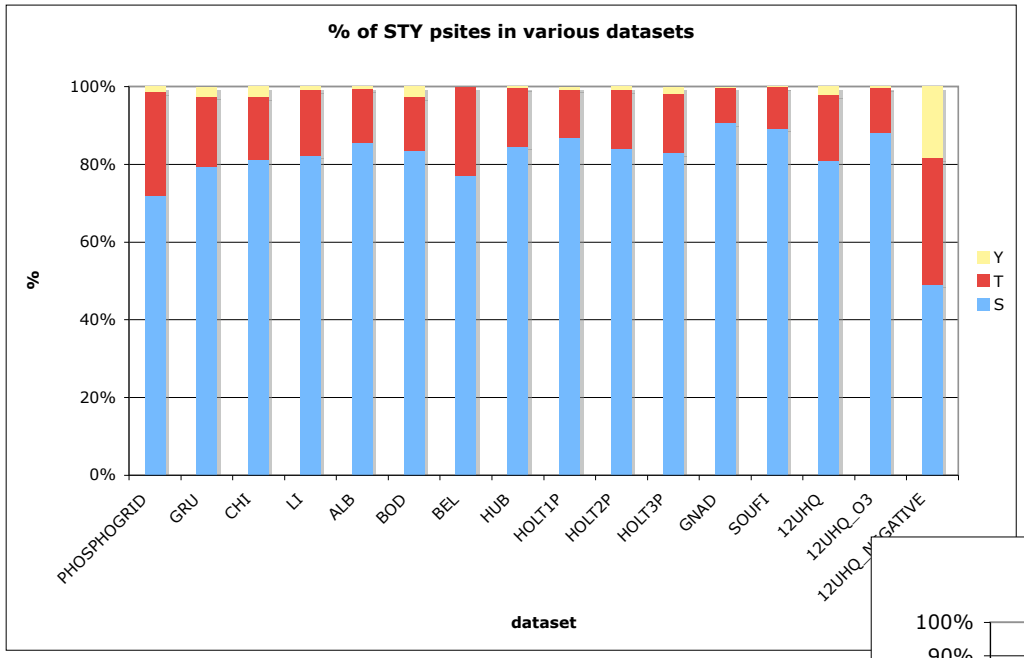
- We suggest that some p-sites are ubiquitously in an 'ON' state (phosphorylated).

- **It may be that the cell keeps a small percentage of the expressed protein molecules of a gene in this phosphorylated state and that this percentage changes according to external stimuli.**

Part C:
Investigating the properties of the phosphoproteome



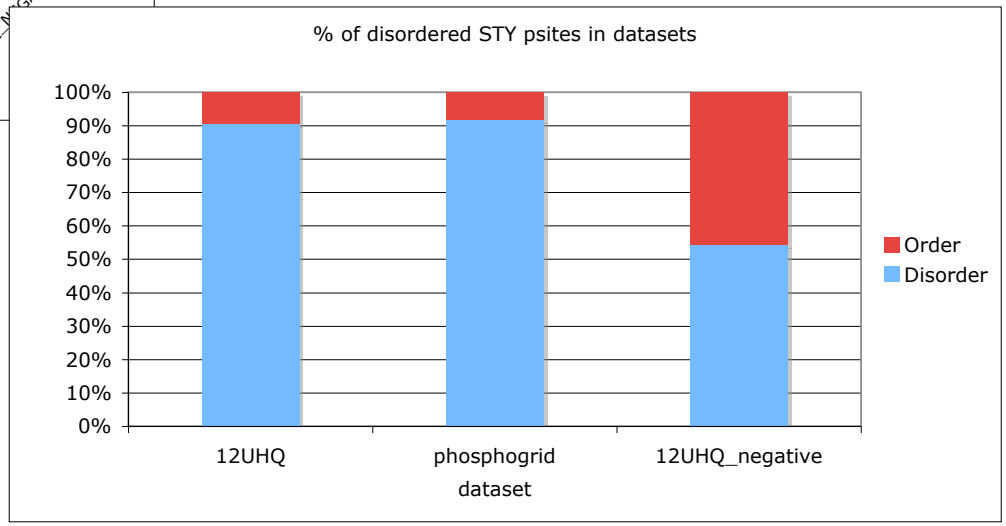
General properties



Serines (81%) >
Threonines (17%) >
Tyrosines (2%).

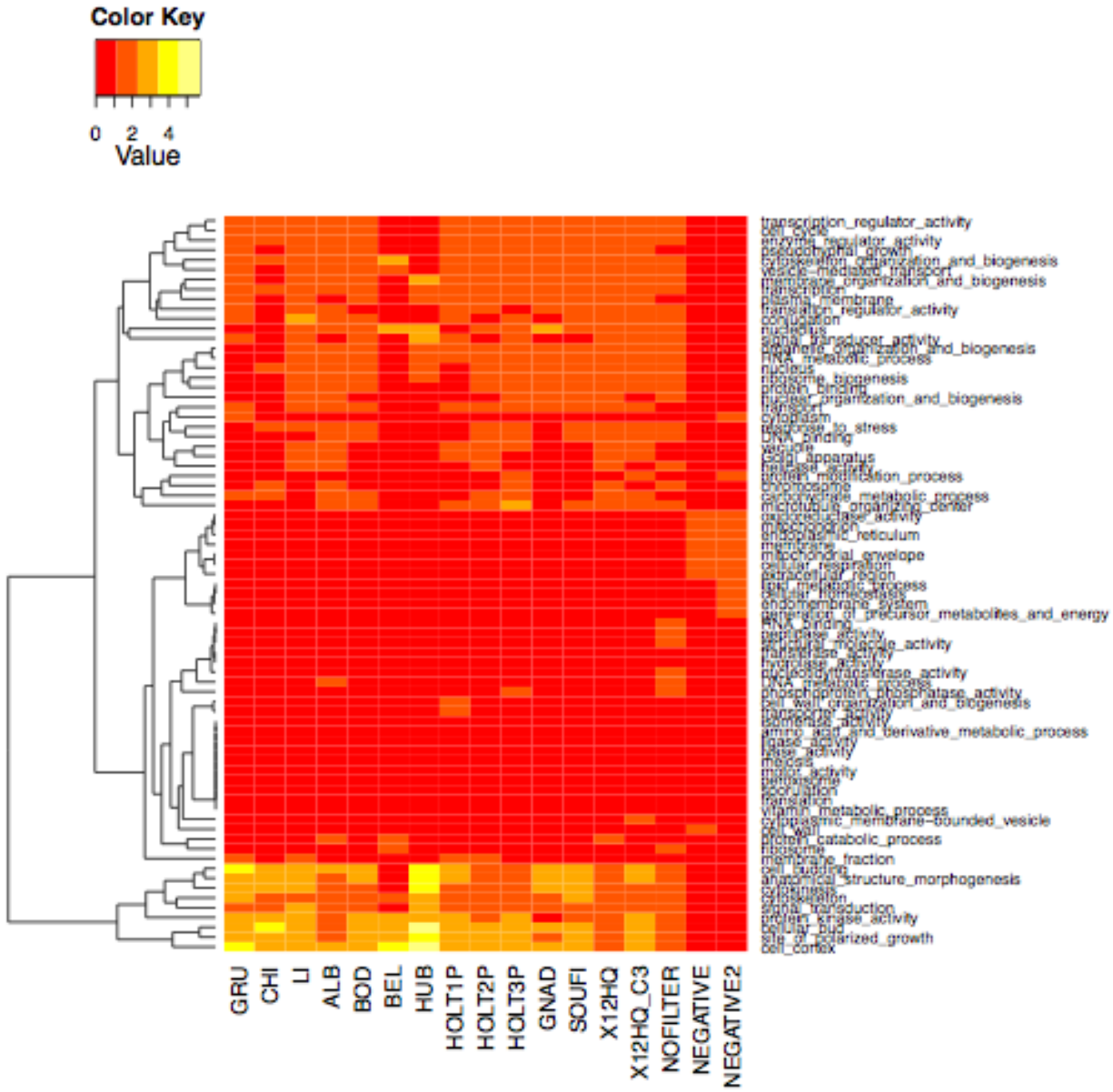
Not an artefact.

P-sites mostly (91%) on disordered regions



17% of 12HQ p-sites and 12% of 12HQ_3 p-sites are found inside or in the vicinity (10 amino acids) of an annotated Pfam domain

GO-Slim

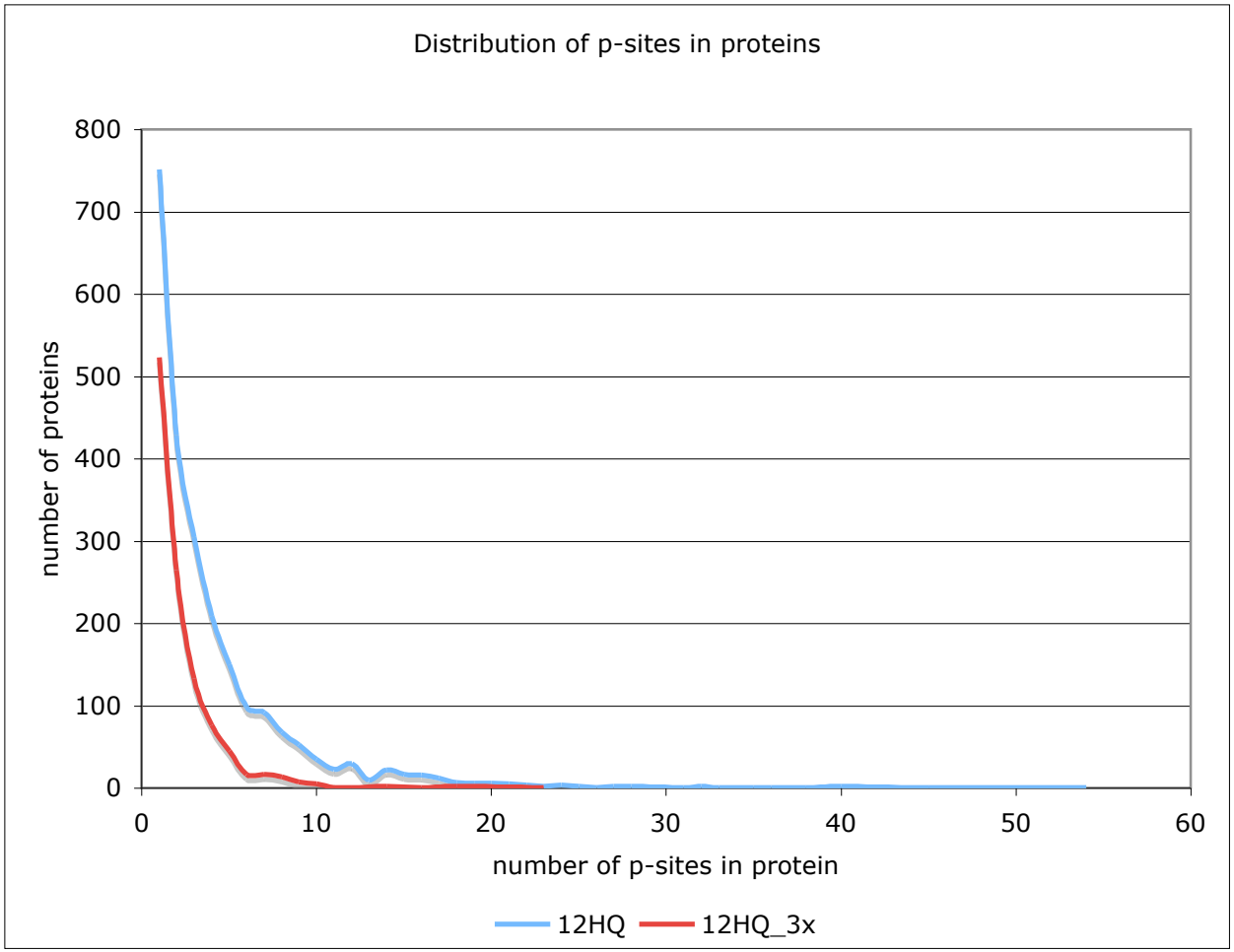


Good reproducibility among experiments

Cell budding

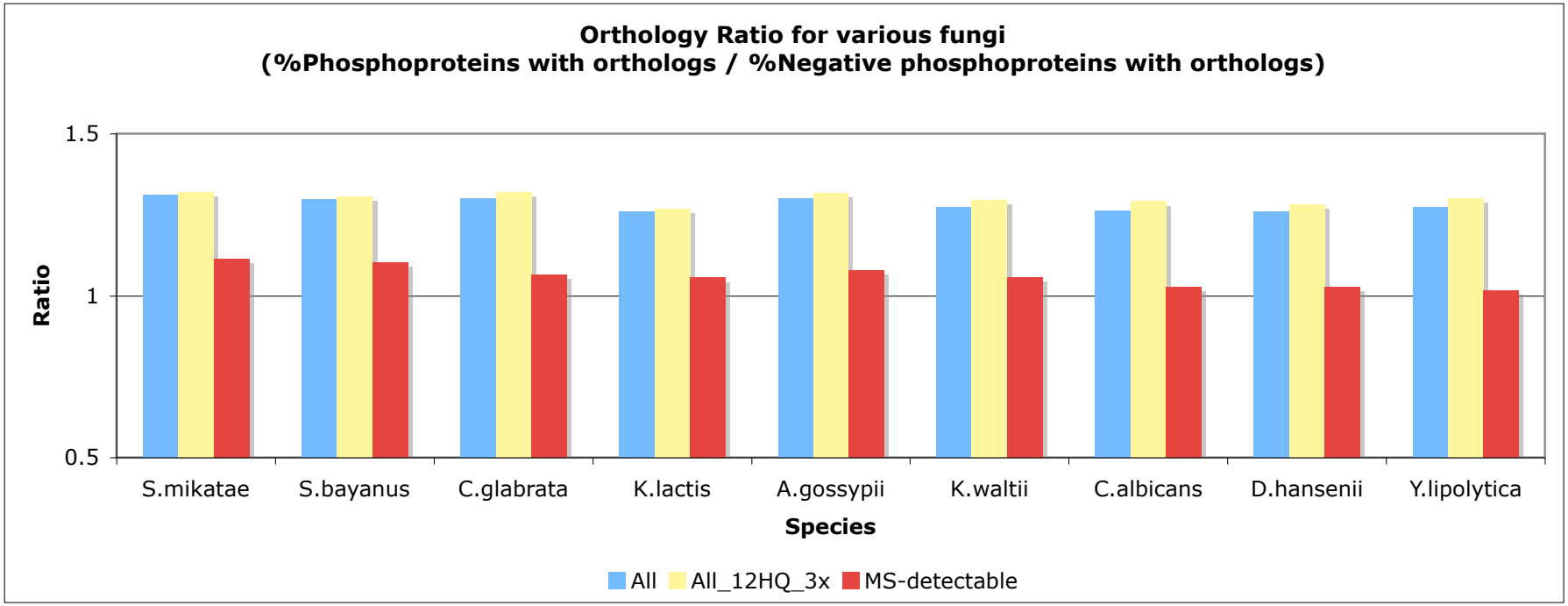
Kinase activity

Distribution of p-sites in proteins



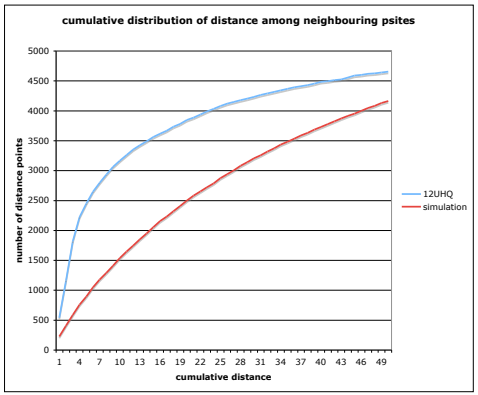
- Similar distribution in other species too.
- The most phosphorylated protein (with 54 p-sites) is Sec16p (YPL085W), which is a coat protein of the COPII vesicle, required for ER transport.

More ancient origin for phosphoproteins

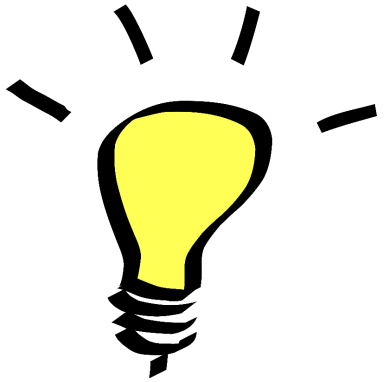


Phosphoproteins vs non-phosphoproteins

- More frequently essential (23% vs 10%) (Chi-squared < $6e^{-23}$).
- WGDs have more psites than singlets
- On average, 50% shorter protein half-life (Wilcoxon p<0.001).
- More frequently ubiquitinated (27% vs 9%) (Chi-squared < $2e^{-16}$).
- On average, 40% more genetic interactions (Wilcoxon p< $2e^{-15}$).
- On average, 48% more protein-protein interactions (Wilcoxon p< $2e^{-13}$).
- 182% longer ID regions (Wilcoxon p=0).
- 38% longer non-ID regions (Wilcoxon p< $2e^{-16}$).
- Weak correlation (Pearson = 0.18) between number of p-sites and kinases targeting the protein
- P-sites tend to cluster



Conclusions



- Yeast Phosphoproteome is incomplete
- The various experiments have similar properties
- Several of the properties that we observed in the current phosphoproteome were also observed correctly in previous and much smaller data sets, with less stringent filtering criteria.
- This high-quality sample is sufficient to accurately reveal the major properties of the entire yeast phosphoproteome.
- Important proteins are more tightly controlled at the post-translational level

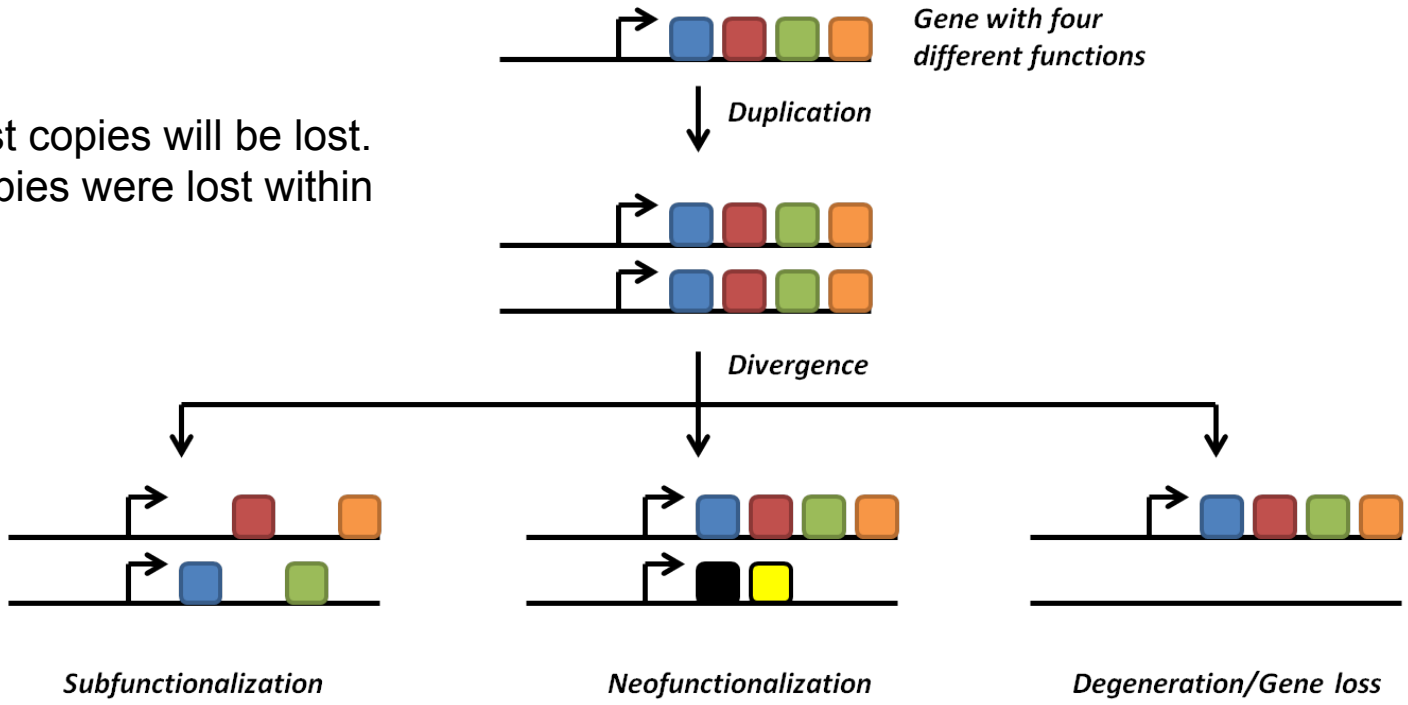
Posttranslational regulation impacts the fate of duplicated genes

Grigoris D. Amoutzias[a,b,1], Ying He[a,b,1], Jonathan Gordon[a,b], Dimitris Mossialos[c], Stephen G. Oliver[d,2], and Yves Van de Peer[a,b,3]

[a]Department of Plant Systems Biology, Flanders Institute for Biotechnology, 9052 Ghent, Belgium; [b]Department of Molecular Genetics, Ghent University, 9052 Ghent, Belgium; [c]Department of Biochemistry and Biotechnology, University of Thessaly, Larissa GR 41221, Greece; and [d]Cambridge Systems Biology Centre and Department of Biochemistry, University of Cambridge, Cambridge CB2 1GA, United Kingdom

Gene duplication

Following WGD, most copies will be lost. In yeast, ~85% of copies were lost within 100 my.



Gene retention could be due to:

- Subfunctionalisation
- Neofunctionalisation
- Dosage balance in protein complexes or regulatory networks
- The need for increased dosage

Whole Genome duplication

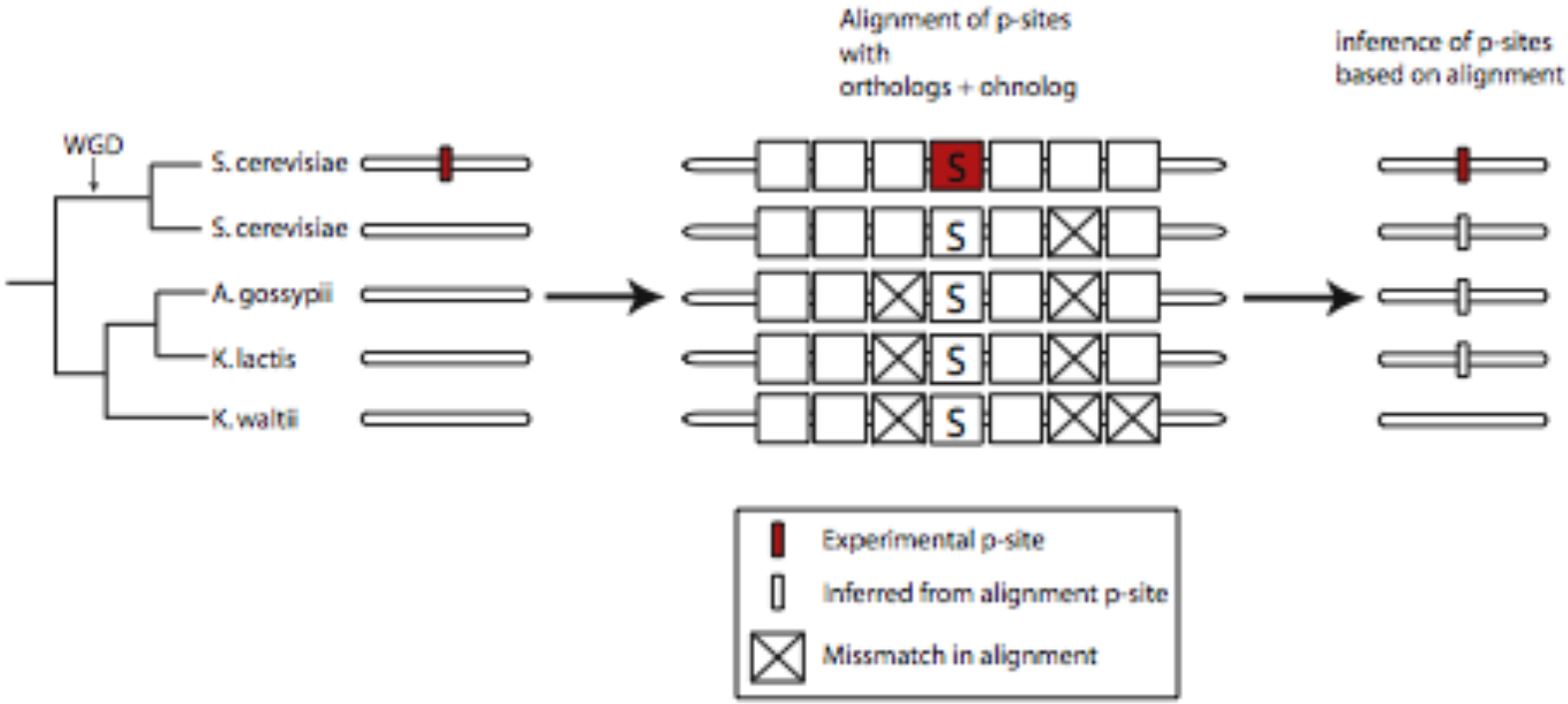
Following WGD, there is a relatively short period of genome instability, extensive gene loss and elevated levels of mutation.

Regulatory networks need to rewire rapidly, to integrate the newly duplicated genes.

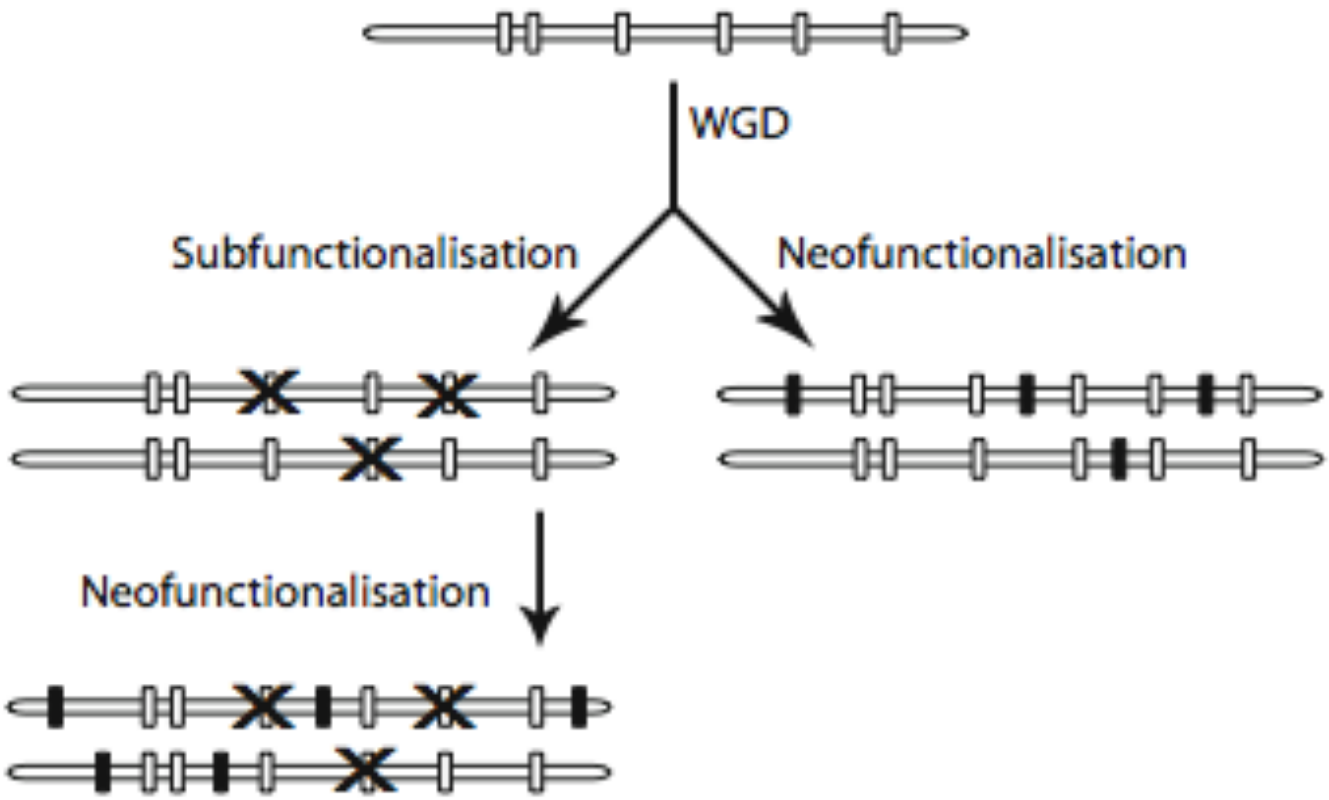
Rapid evolution has been observed at the level of transcription of duplicated genes, by mutations in short transcription factor binding motifs.

Nevertheless, the effectors of gene action are the proteins.
Rapid changes could occur at the post-translational level of regulation too.

Inferring the ancestral phosphorylation state



Inferring the ancestral phosphorylation state



Conclusions

- WGD phosphoproteins have on average, more p-sites than RSS proteins.
- This is a general trend, found in many, though not all gene categories.
- We controlled for potential biases stemming from protein abundance, coverage of experiments, dosage balance hypothesis.
- Ancestral proteins that were later retained as duplicates already had more psites.
- Subfunctionalisation and neofunctionalisation could be some of the reasons behind gene retention.
- WGD proteins generally have tighter post-translational regulation (ubiquitination, half-lives) than RSS proteins.
- This trend is observed for Single-gene duplicates too.
- This trend seems to hold for other species too.

G3 Genes | Genomes | Genetics

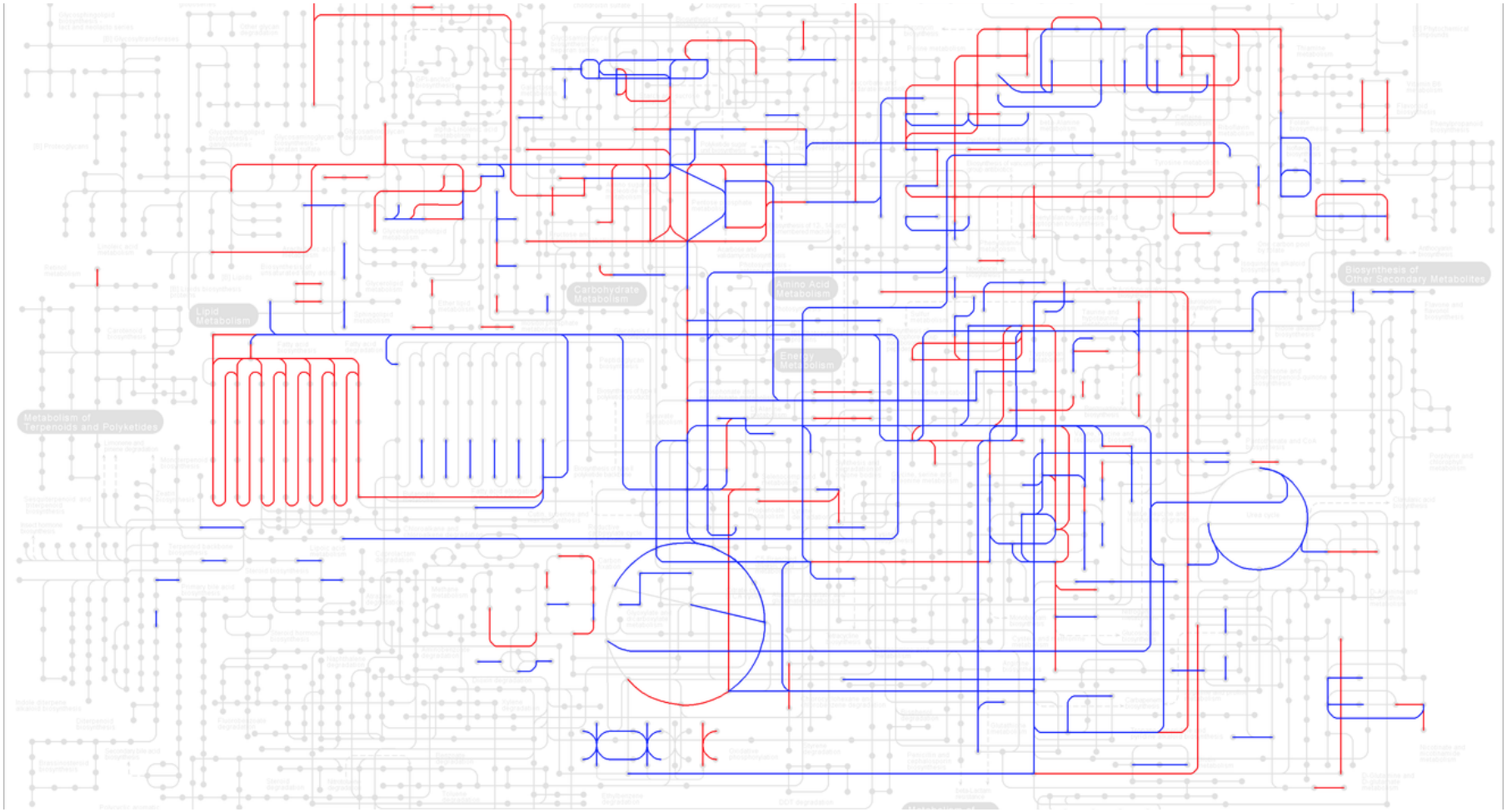
INVESTIGATION

The Pivotal Role of Protein Phosphorylation in the Control of Yeast Central Metabolism

Panayotis Vlastaridis,* Athanasios Papakyriakou,*,† Anargyros Chaliotis,* Efstratios Stratikos,† Stephen G. Oliver,‡,§,1 and Grigorios D. Amoutzias*,1

*Bioinformatics Laboratory, Department of Biochemistry & Biotechnology, University of Thessaly, Biopolis, Larisa 41500, Greece, †National Centre for Scientific Research Demokritos, Agia Paraskevi 15341, Greece, and ‡Cambridge Systems Biology Centre and §Department of Biochemistry, University of Cambridge, CB2 1GA, UK

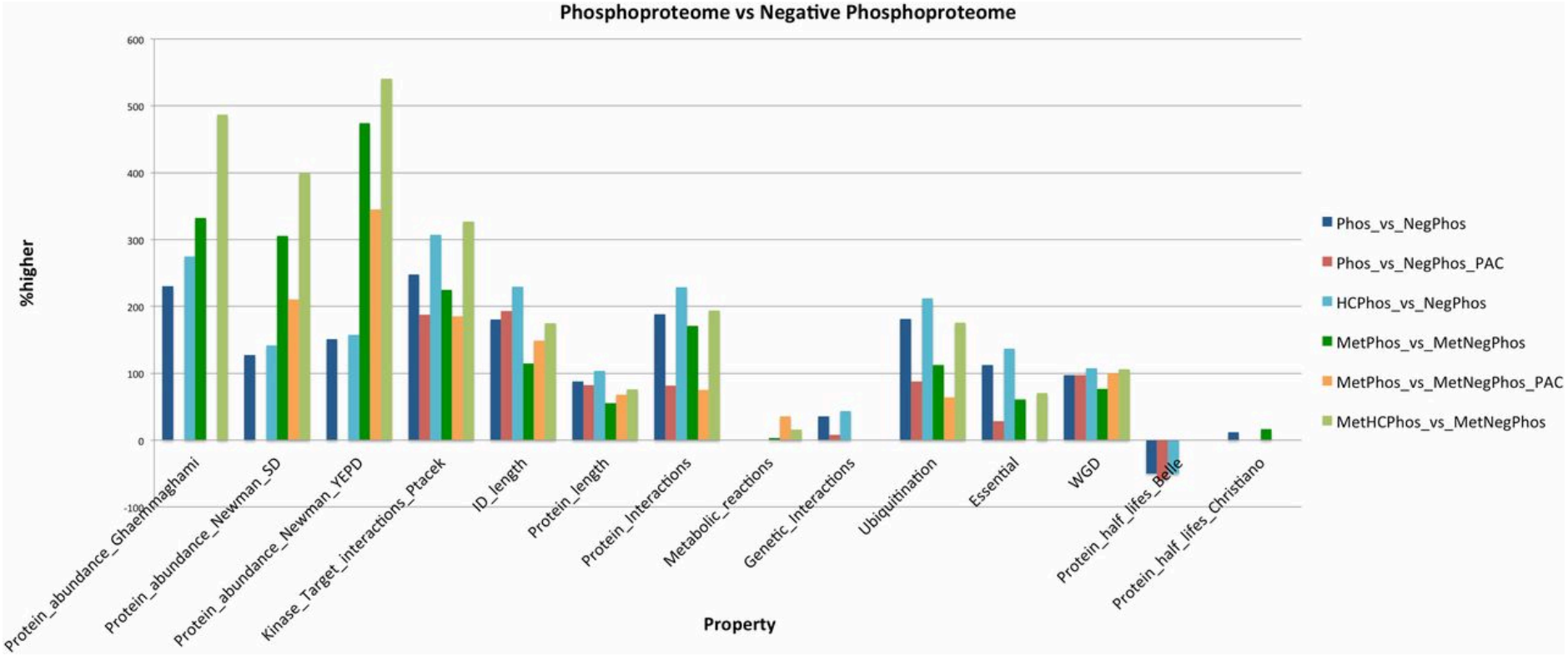
Half of the metabolic proteins are regulated by phosphorylation



Important enzymes are usually phosphorylated

	Phospho-metabolic vs rest metabolic proteins
Protein abundance	137%-326% higher
Intrinsically disordered regions	90%-117% longer
Protein-protein interactions	86%-131% more
Kinase-target interactions	171%-178% more
Essential	17-18% vs 10-12%
Ubiquitinated	41-53% vs 23-25%
Whole genome duplicates	28-32% vs 18-20%

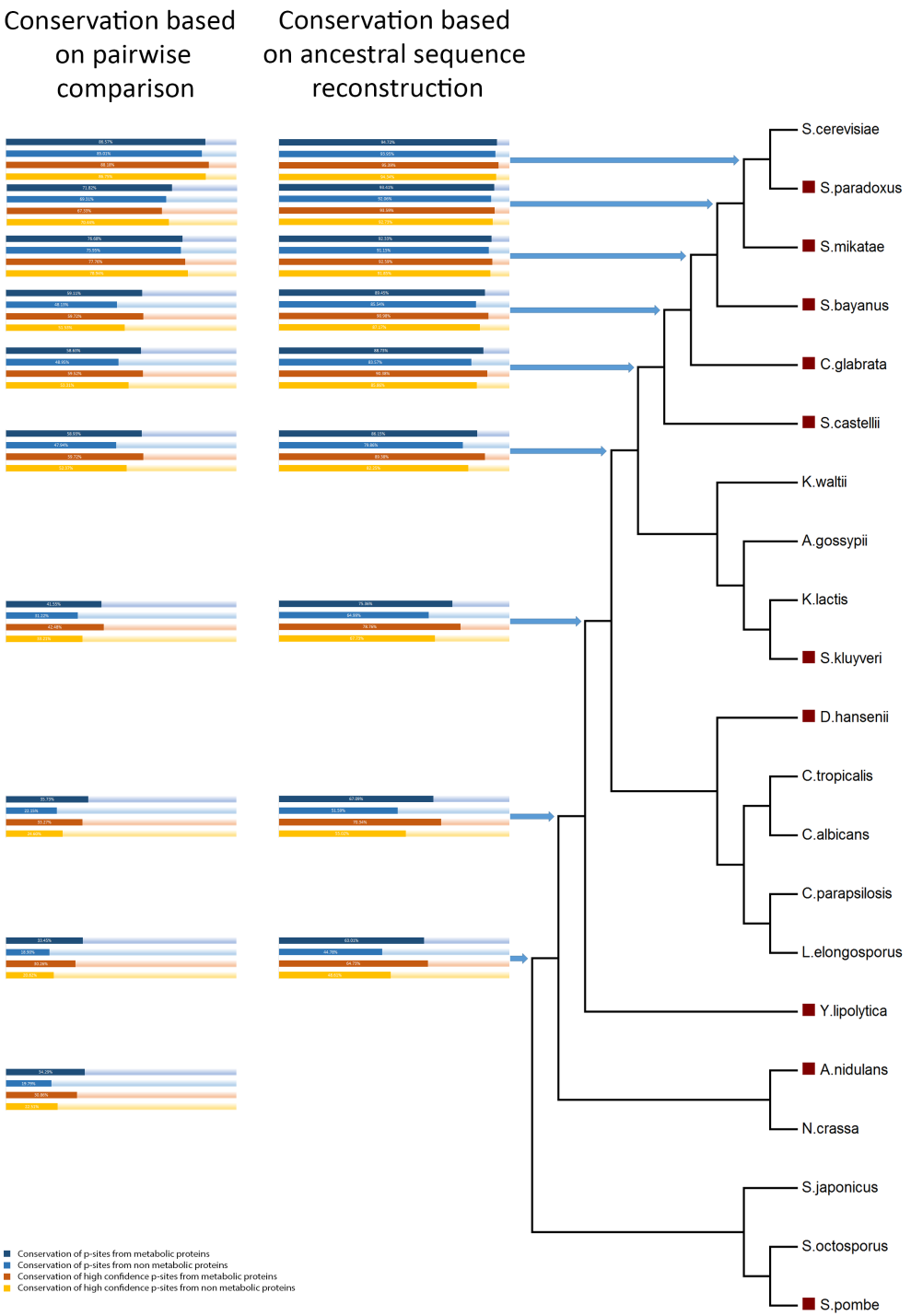
The general properties of the phosphoproteome, compared to the negative phosphoproteome.



Panayotis Vlastaridis et al. G3 2017;7:1239-1249

Metabolic P-sites are more conserved

Prediction in other species



How much can we learn from other species: Comparative Phosphoproteomics

- In yeast
  - 692 psites of 431 orfs have a conserved and identified p-site in C.albicans
  - 477 p-sites of 296 orfs have a conserved and identified p-site in human.

Comparative phosphoproteomics could increase the yeast phosphoproteome by 15%.

P-sites evolve fast

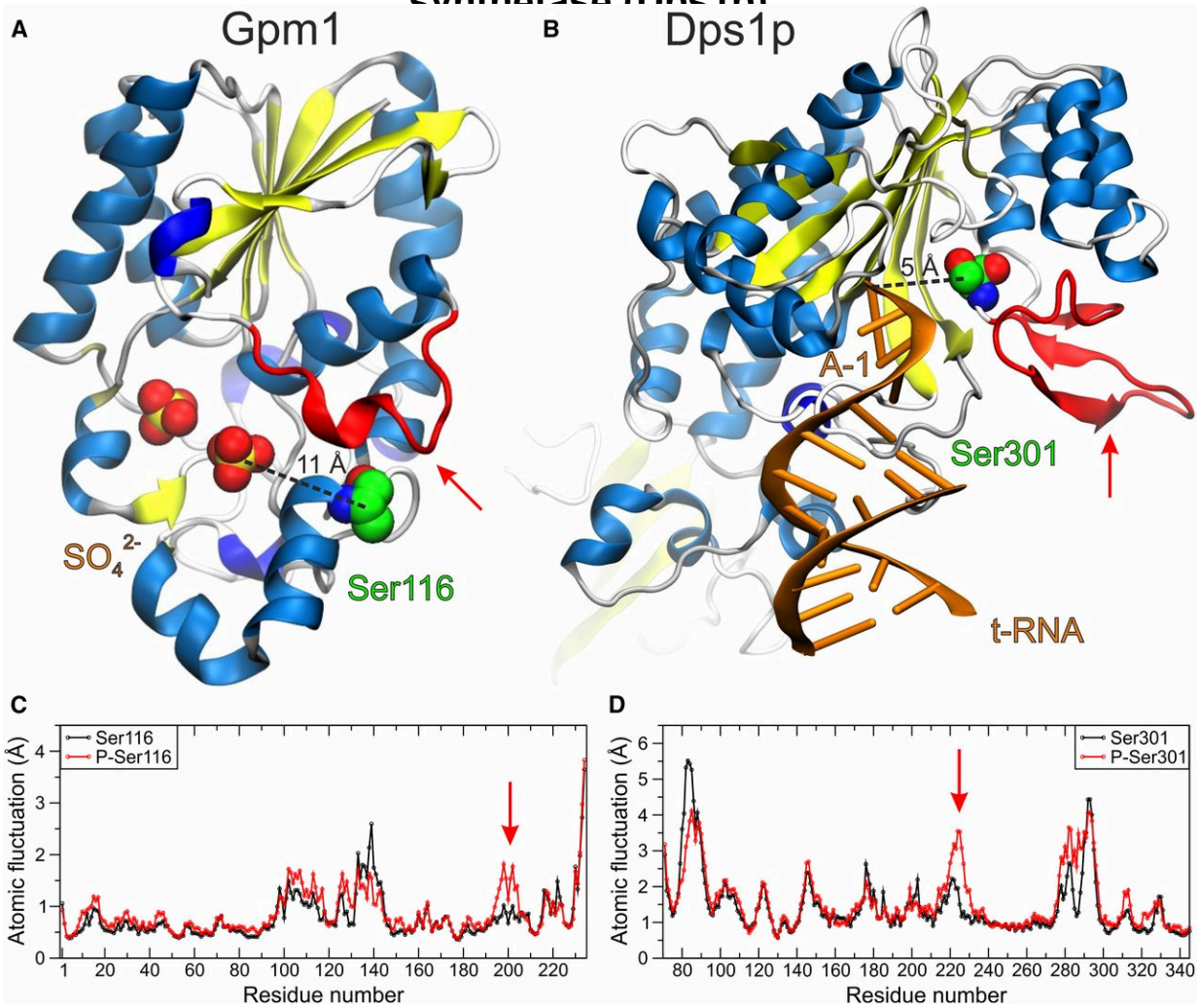
Yeast	**M**	**A**	**K**	**P**	**S**	**R**	**L**	**I**	**T**	**K**	**P**
C. albicans	**M**	**L**	**K**	**P**	**S**	**R**	**–**	**L**	**T**	**K**	**P**

Could phosphorylation be used in biotechnological applications?

Gene Group	P-sites/proteins
Essential	3025/576
Metabolism essential	339/71
Biotechnological Phenotypes	2363/408

Phenotype_terms	Psites/prots
chemical compound excretion: increased	1497/248
fermentative growth: increased	7_/_3
fermentative metabolism: increased	85/10
growth in exponential phase: increased	73/8
nutrient uptake/utilization: increased	124/20
respiratory growth: increased	416/75
respiratory metabolism: increased	331/61
utilization of carbon source: increased	36/8
vegetative growth: increased	8_/_5
viability: increased	67/17
ALL_RELATED_Phenotypes	2363/408

Molecular representations of two p-sites examined with molecular dynamic simulations in (A) the yeast phosphoglycerate mutase (Gpm1p) and (B) aspartyl-tRNA (transfer RNA) synthetase (Dps1p)



Panayotis Vlastaridis et al. G3 2017;7:1239-1249

Acknowledgements

Dept. of Biochemistry & Biotechnology,

University of Thessaly, Greece

- Panayotis Vlastaridis
- Pelagia Kyriakidou
- Anargyros Chaliotis

Department of Biochemistry,

University of Cambrdige, UK

- Steve Oliver
- Kathryn Lilley

NCSR Dimokritos

- Stratos Stratikos
- Thanos Papakyriakou

VIB, Plant Systems Biology,

UGent, Belgium.

- Yves Van de Peer
- Ying He

European Union
European Social Fund

OPERATIONAL PROGRAMME
EDUCATION AND LIFELONG LEARNING
investing in knowledge society
MINISTRY OF EDUCATION & RELIGIOUS AFFAIRS
MANAGING AUTHORITY

NSRF 2007-2013
programme for development
EUROPEAN SOCIAL FUND

Co-financed by Greece and the European Union

EMBO

This project is implemented under the "ARISTEIA II" Action of the "OPERATIONAL PROGRAMME EDUCATION AND LIFELONG LEARNING" and is co-funded by the European Social Fund (ESF) and National Resources.

OXFORD

(GIGA)n SCIENCE

GigaScience, 6, 2017, 1–11

doi: 10.1093/gigascience/giw015
Advance Access Publication Date: 7 January 2017
Research

RESEARCH

Estimating the total number of phosphoproteins and phosphorylation sites in eukaryotic proteomes

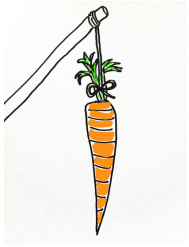
Panayotis Vlastaridis[1], Pelagia Kyriakidou[1], Anargyros Chaliotis[1], Yves Van de Peer[2,3,4], Stephen G. Oliver[5] and Grigoris D. Amoutzias[1,*]

[1]Bioinformatics Laboratory, Department of Biochemistry and Biotechnology, University of Thessaly, Larisa, 41500, Greece, [2]Department of Plant Systems Biology, VIB and Department of Plant Biotechnology and Bioinformatics, Ghent University, B-9052 Ghent, Belgium, [3]Bioinformatics Institute Ghent, Technologiepark 927, B-9052 Ghent, Belgium, [4]Department of Genetics, Genomics Research Institute, University of Pretoria, Pretoria 0028, South Africa, and [5]Cambridge Systems Biology Centre & Department of Biochemistry, University of Cambridge, Cambridge CB2 1GA, UK.

*Correspondence: amoutzias@bio.uth.gr

- All data available
- Video of curve-fitting in excel

Motivation



HTP phosphoproteomics has revolutionized the field and provided unique insight in a whole level of cell regulation.

But we are still discovering new phosphorylation sites.

We need to have an estimate of the total size, to know where we are and where we need to go.

Past Suggestions:

1/3 – 2/3 of the proteome

For human p-sites:

- 57K
- 500K
- 700K
- 1M

Datasets used

Scanned >1000 publications

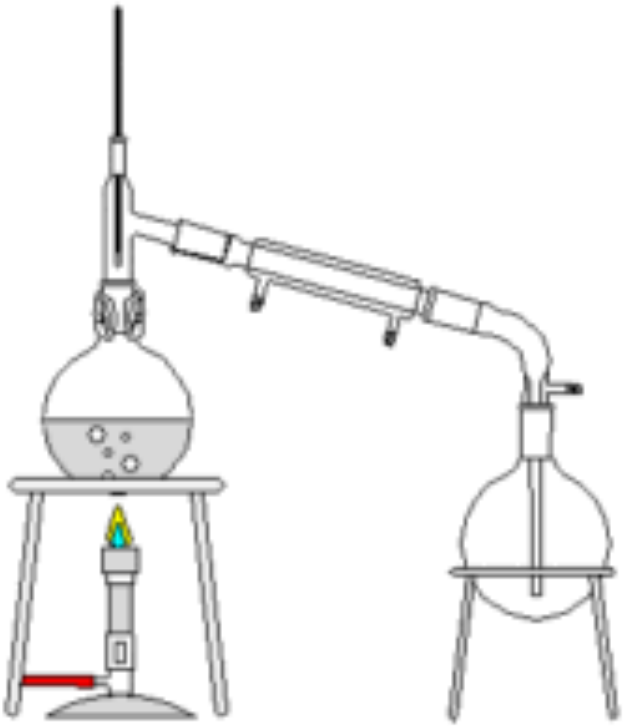
187 high-throughput phosphoproteomic datasets were filtered, compiled and studied along with two low-throughput compendia.

- Human:97
- Mouse:42
- Yeast:20
- Arabidopsis:28
- PhosphoGrid2 (for yeast LTP)
- Phosphosite + (for human and mouse LTP)

Filtering out “noisy” p-sites

99% correct peptide identification
99% correct p-site localization

Very stringent criteria for individual analyses
Needed when compiling compendiums



Estimation methods

Capture-Recapture: Established method in Ecology and Epidemiology. Based on overlap among the various experiments

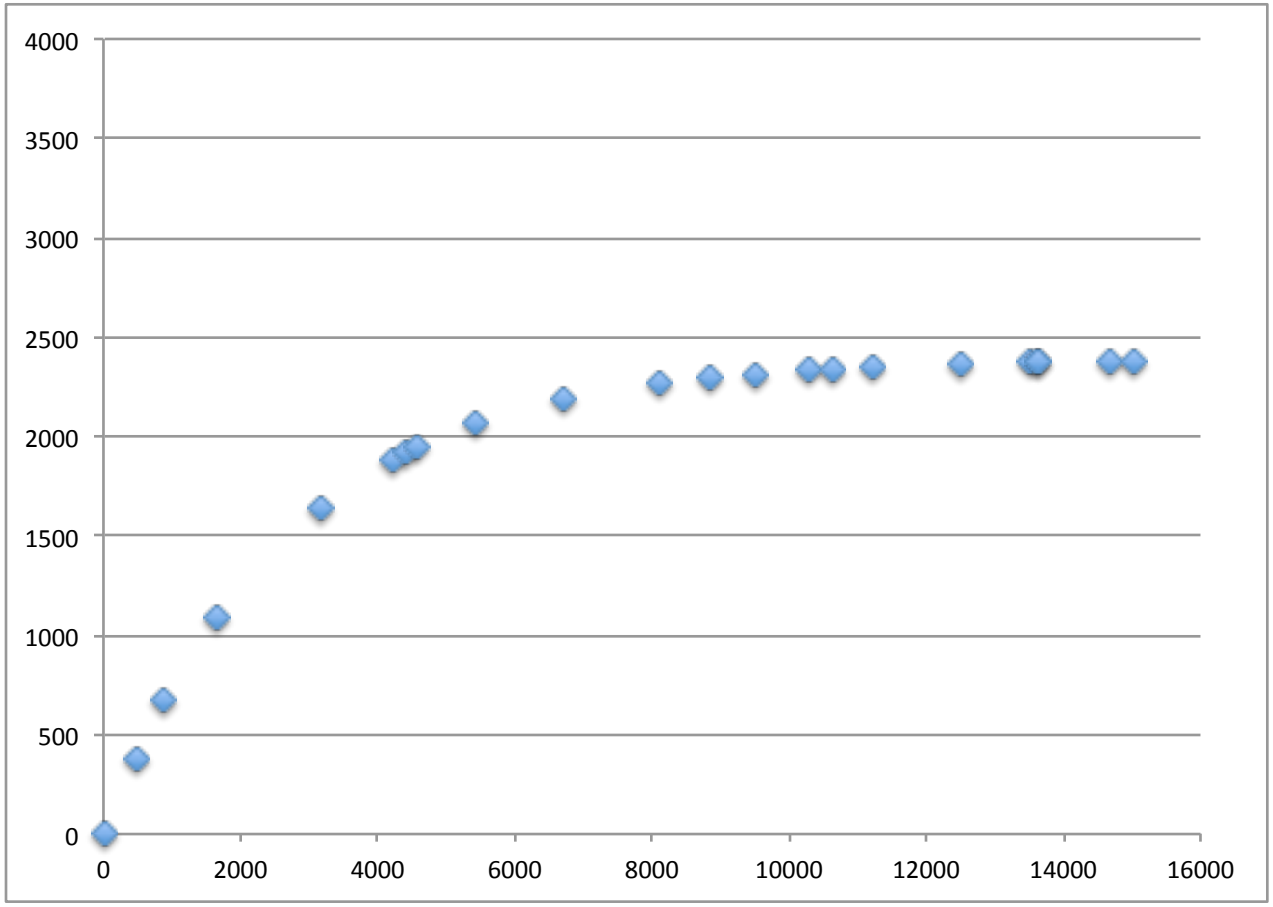
Curve-fitting the saturation curve of cumulative redundant vs. cumulative non-redundant phosphoproteins/p-sites.

- Modeled by exponential recovery function.
- Can also model different noise levels

Estimates were also adjusted for different levels of noise (1,5,10%) within the individual datasets and also permutated the data to observe robustness of conclusions

The saturation curve of a compendium

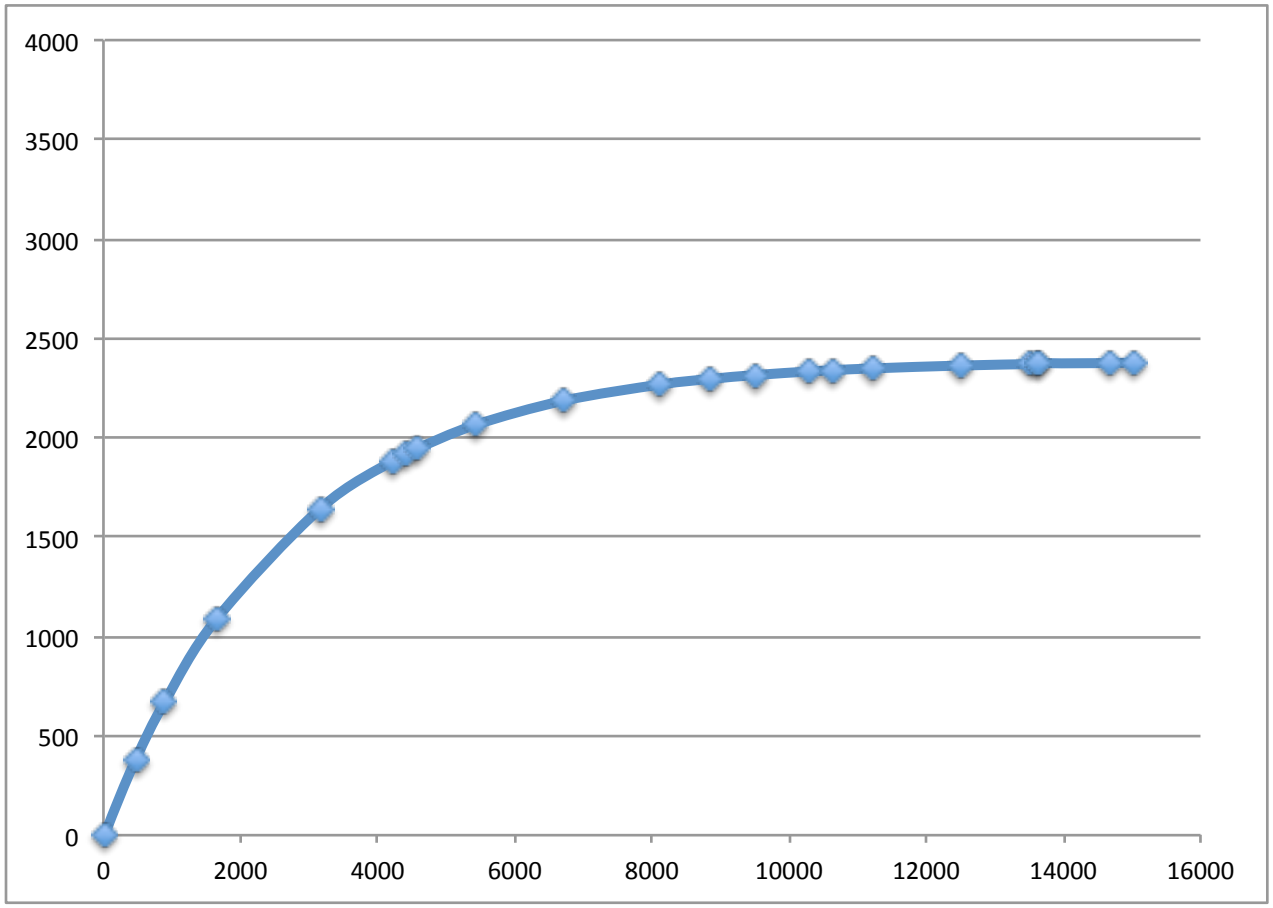
Unique proteins



Cumulative number of proteins

The saturation curve: Exponential recovery

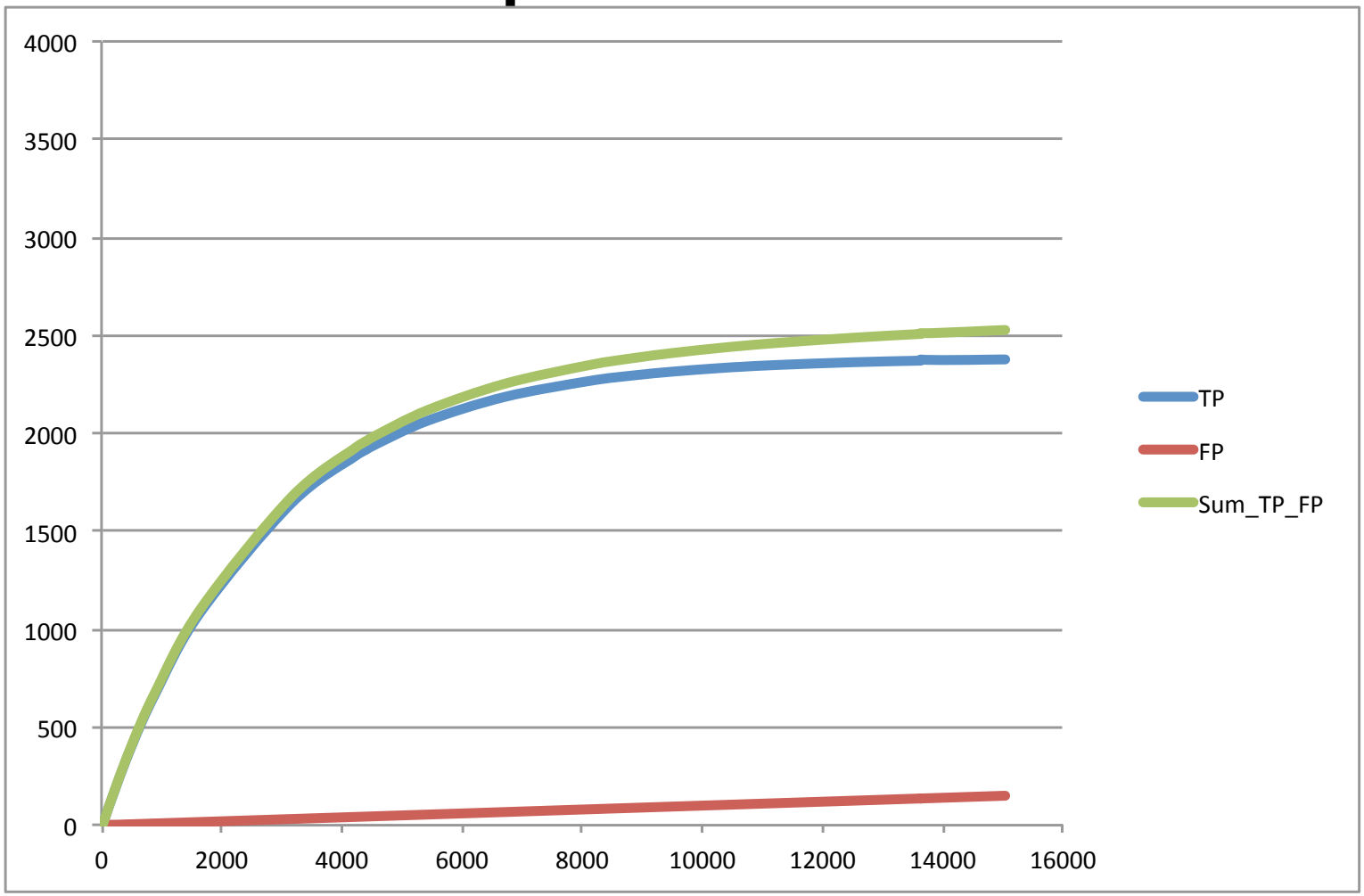
Unique proteins



Cumulative number of proteins

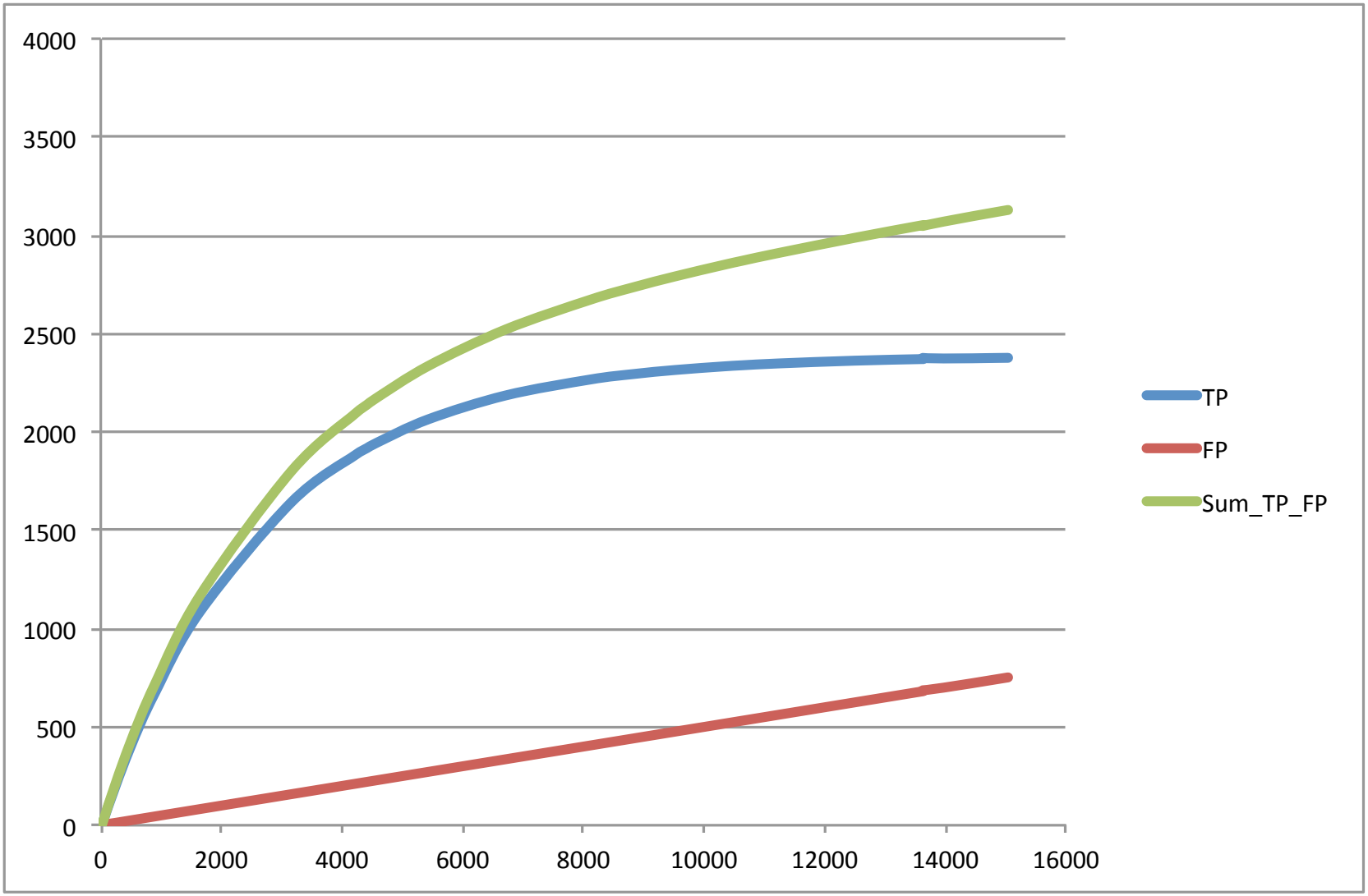
$y = a(1-e^{(-x/b)})$

Adding Noise:1% average noise per dataset

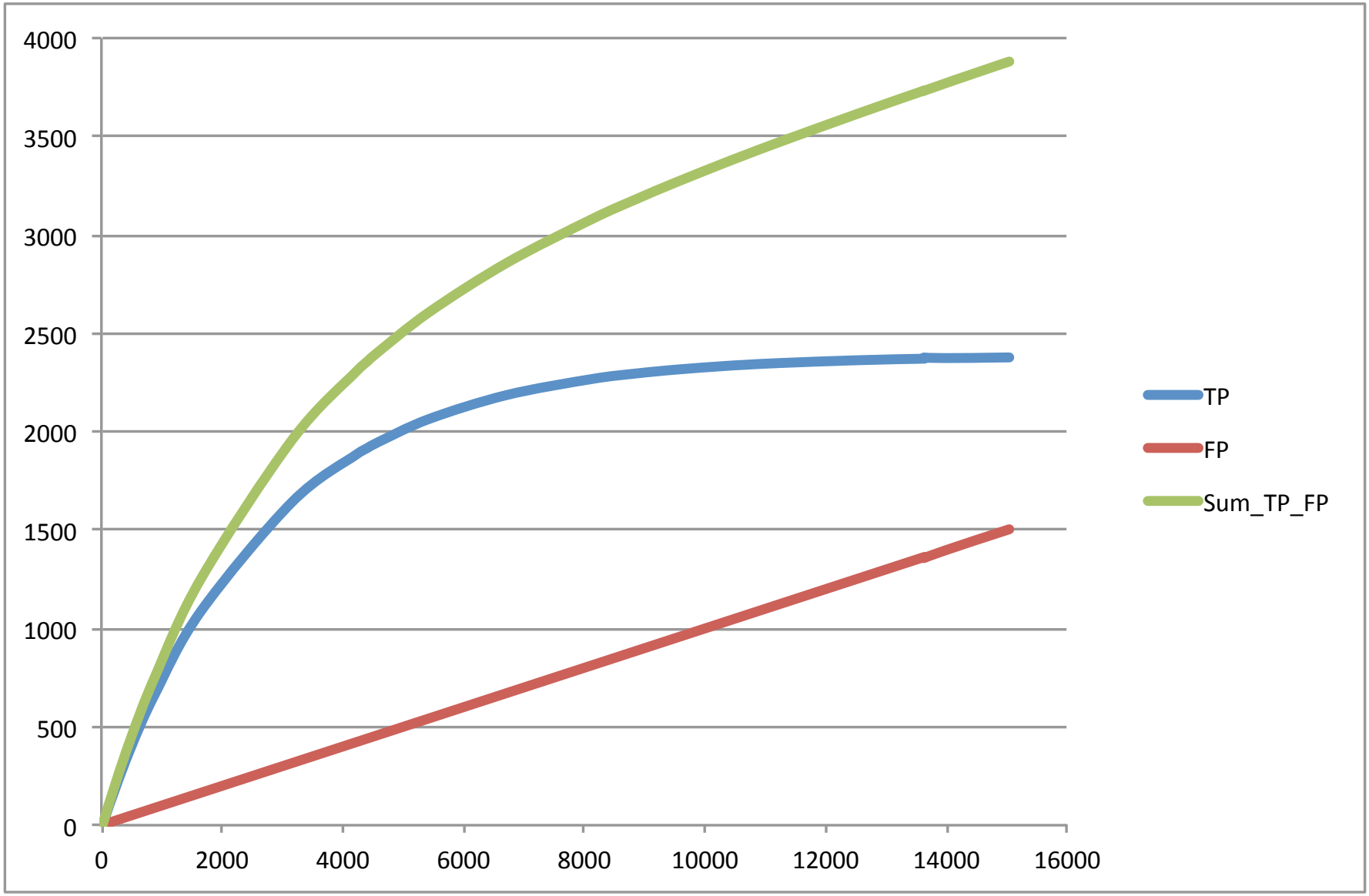


$$y = a(1-e^{(-x/b)}) + 0.01x$$

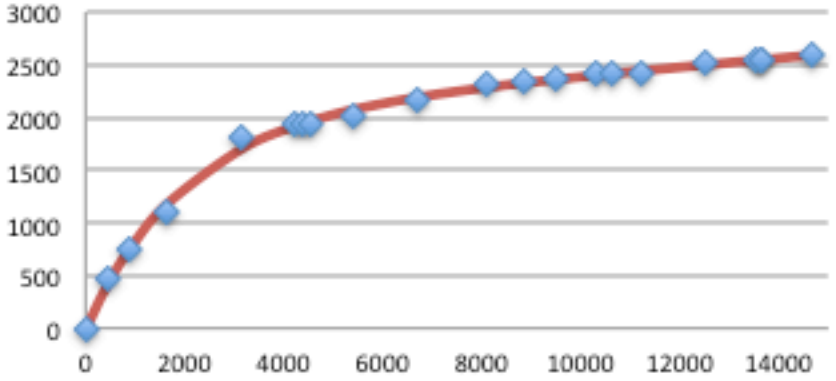
5% average noise



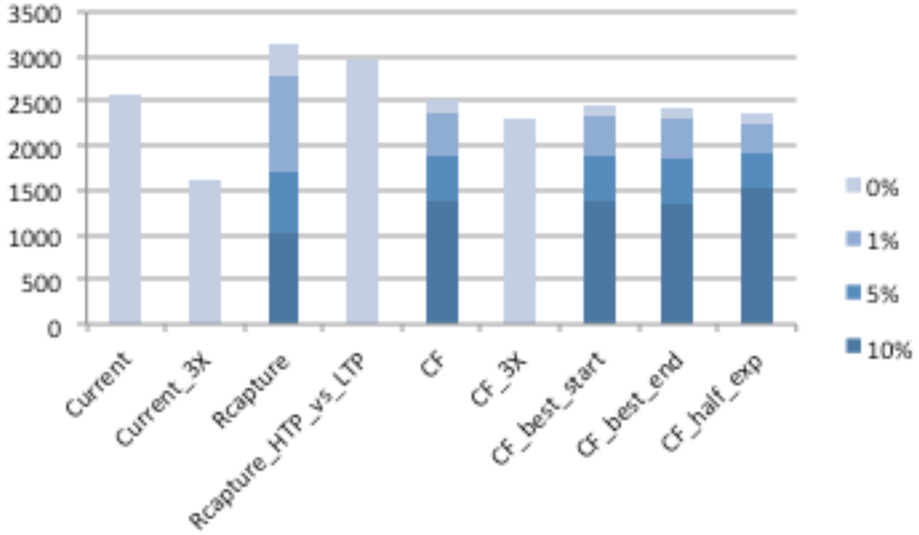
10% average noise



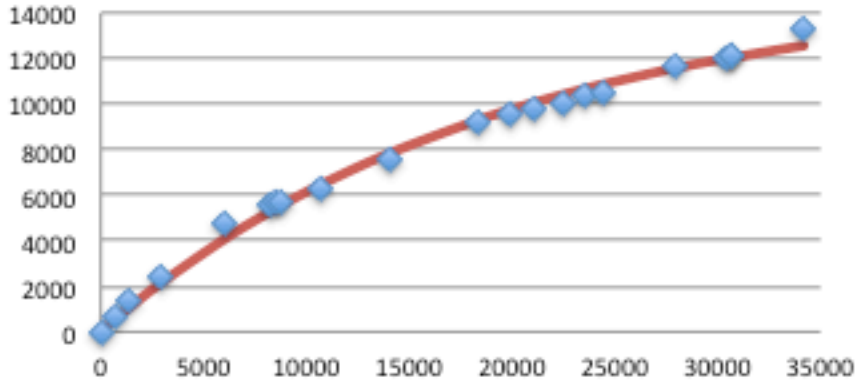
A) Saturation curve of yeast phosphoproteins



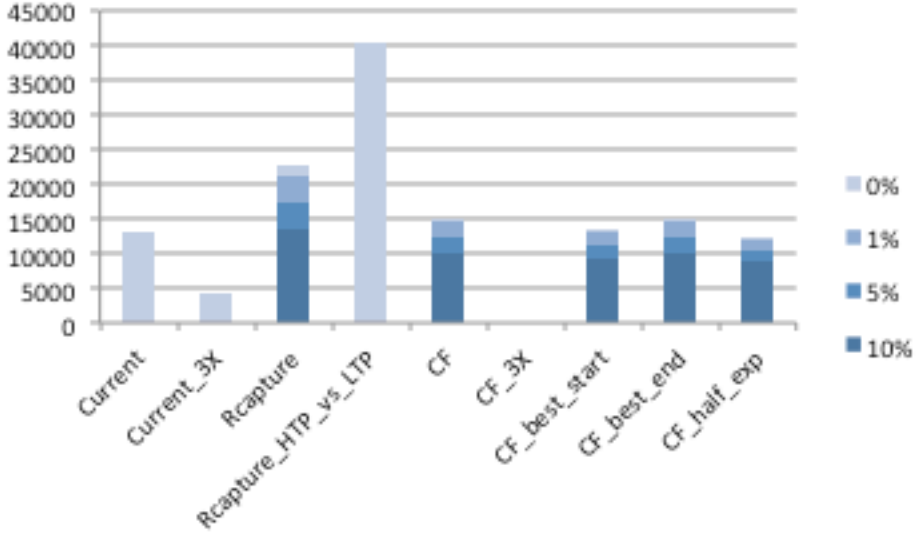
B) Estimation of total yeast phosphoproteins



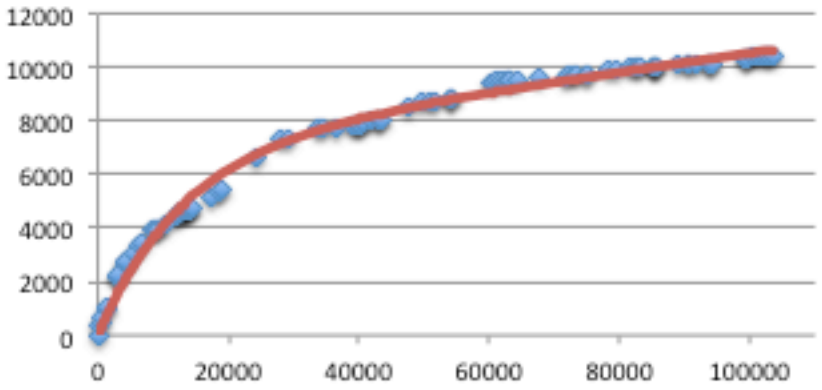
C) Saturation curve of yeast p-sites



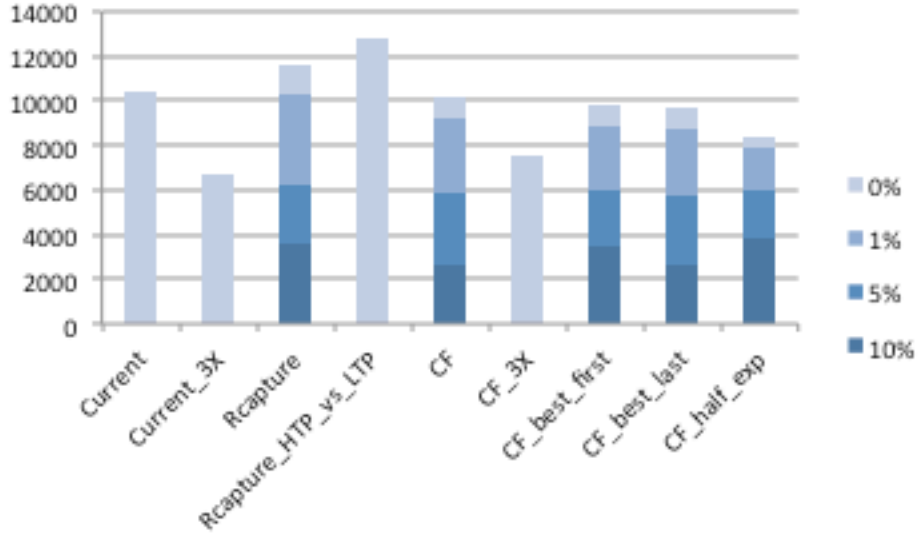
D) Estimation of total yeast p-sites



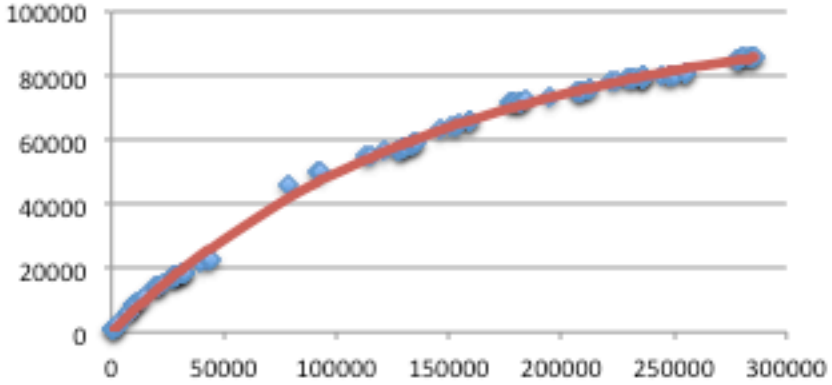
A) Saturation curve of human phosphoproteins



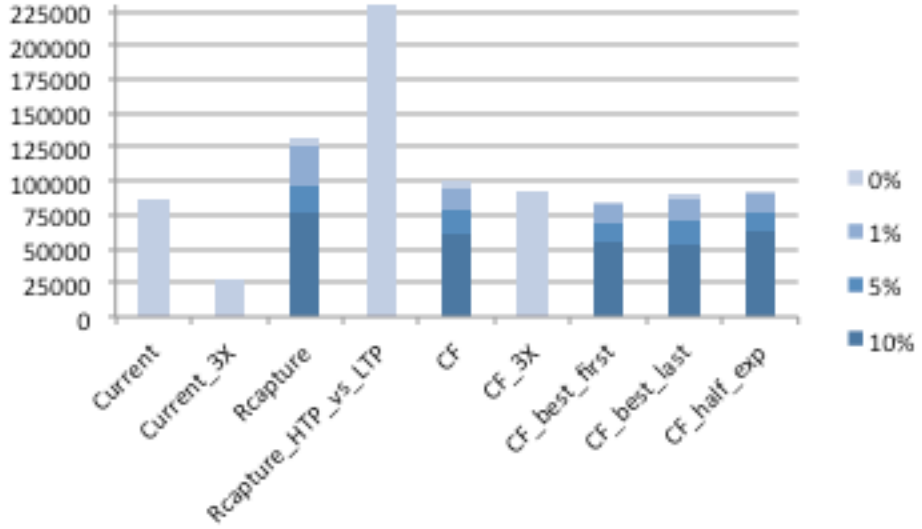
B) Estimation of total human phosphoproteins



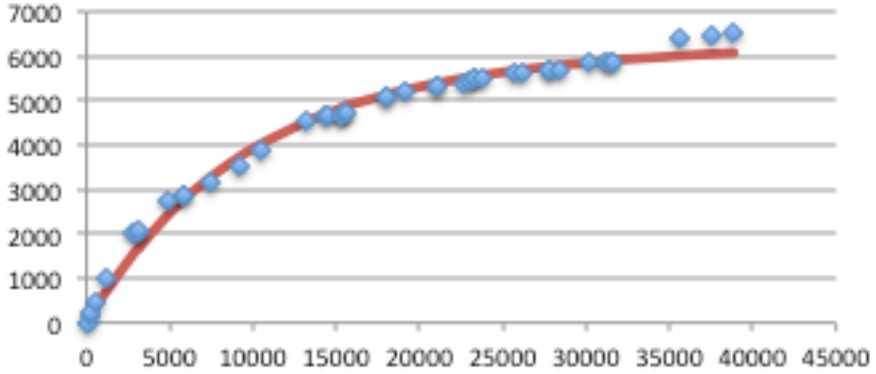
C) Saturation curve of human p-sites



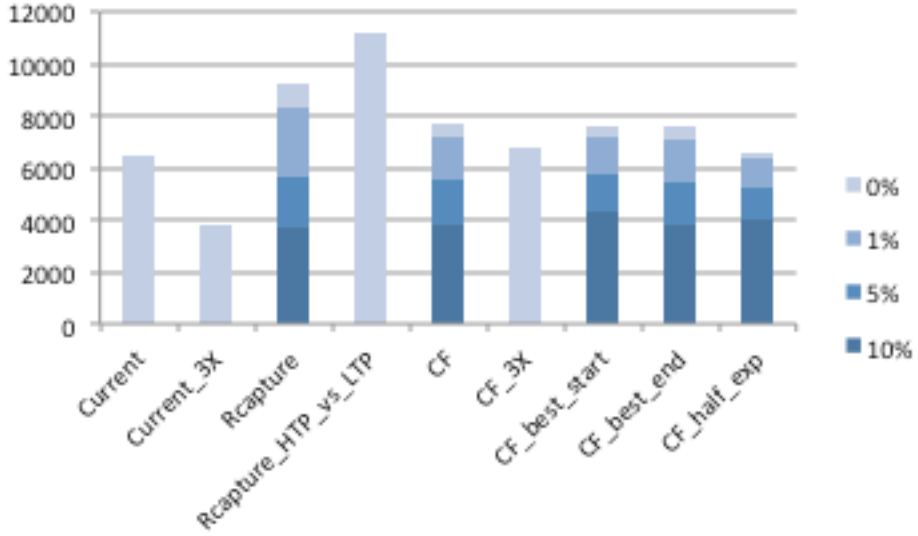
D) Estimation of total human p-sites



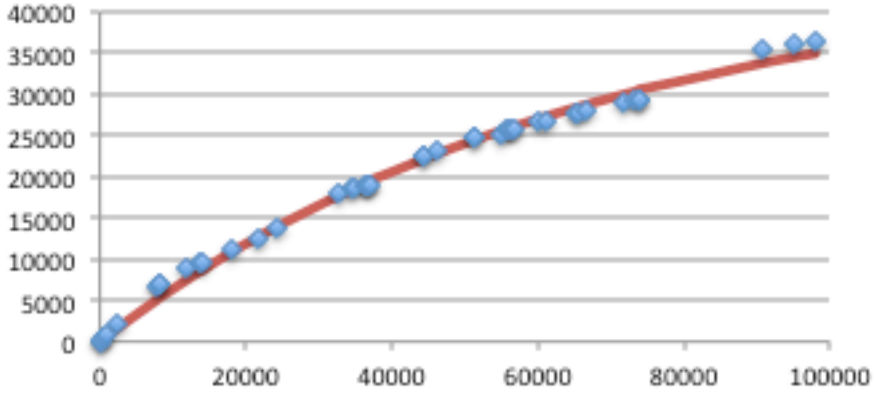
A) Saturation curve of mouse phosphoproteins



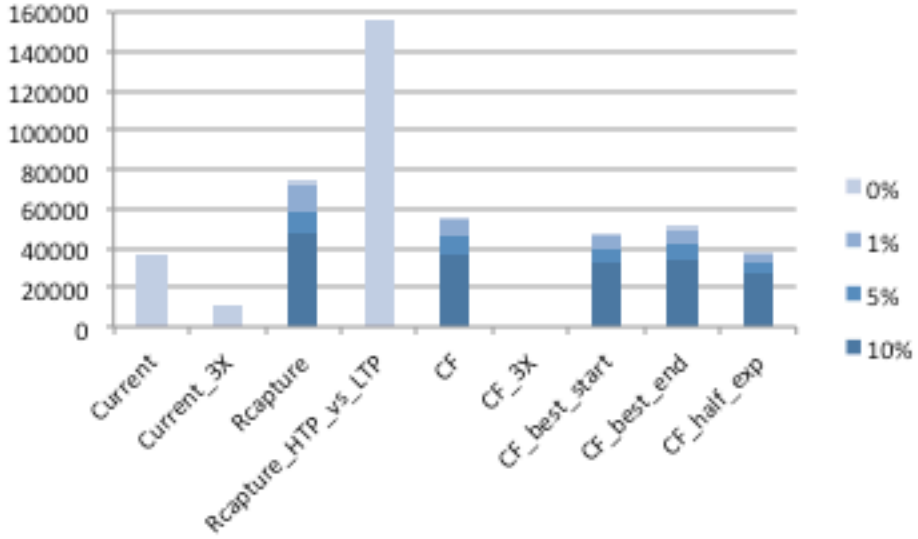
B) Estimation of total mouse phosphoproteins



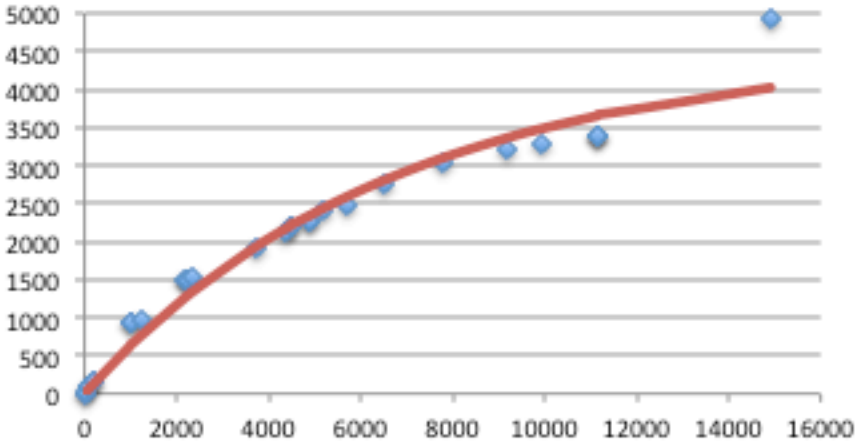
C) Saturation curve of mouse p-sites



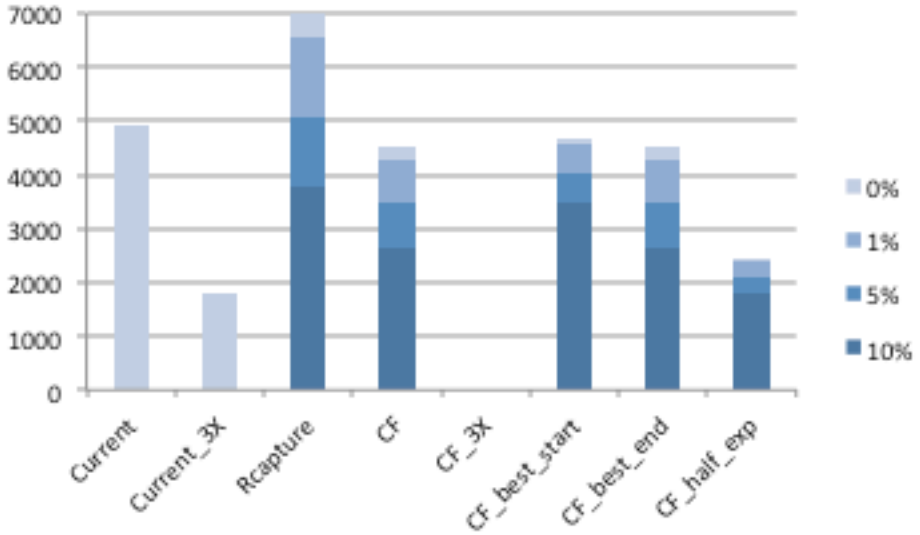
D) Estimation of total mouse p-sites



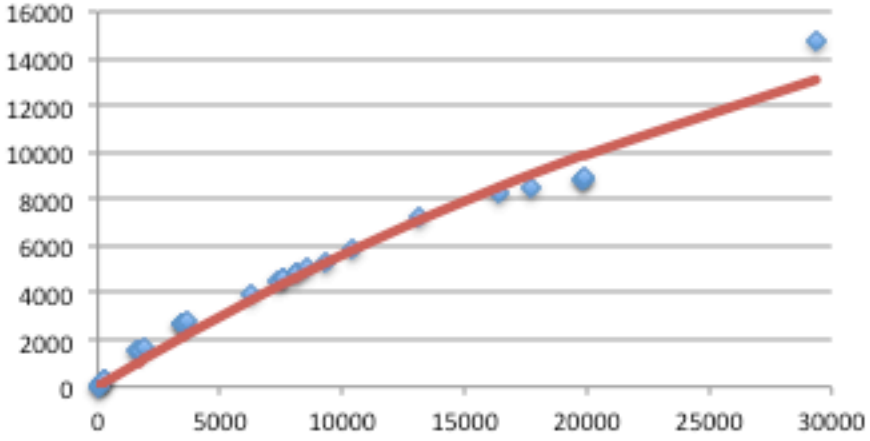
A) Saturation curve of *Arabidopsis* phosphoproteins



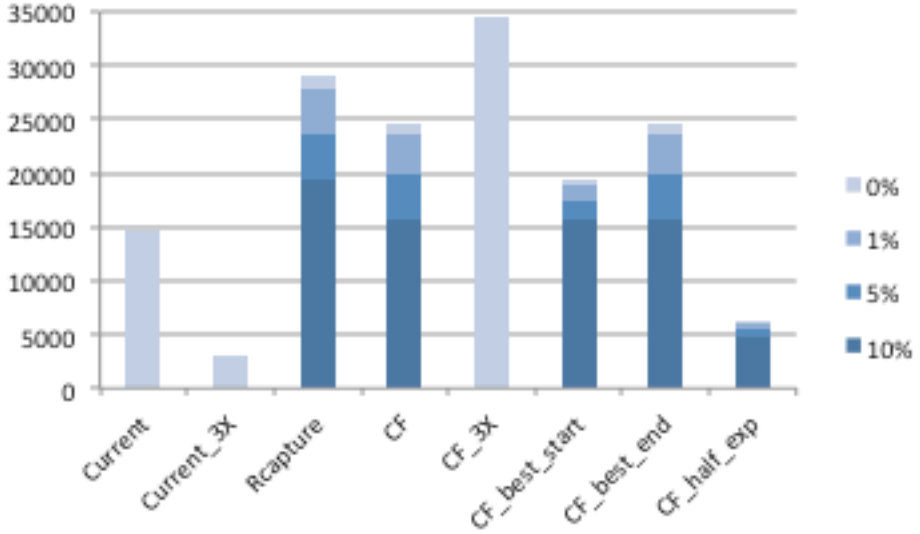
B) Estimation of total *Arabidopsis* phosphoproteins



C) Saturation curve of *Arabidopsis* p-sites



D) Estimation of total *Arabidopsis* p-sites



Conclusions

- Most of the phosphoproteins have been discovered for human, mouse and yeast, while the dataset for *Arabidopsis* is still far from complete.
- The datasets for p-sites are not as close to saturation as those for phosphoproteins.
- Integration of the low-throughput data suggests that current high-throughput phosphoproteomics is capable of capturing 70-95% of total phosphoproteins & 40-60% of total p-sites.
- More datasets needed to provide more accurate estimates in the future.
- Capture-Recapture and Curve-fitting should be used to estimate completeness of experimental replicates