Evolution of Protein Post-translational Regulation

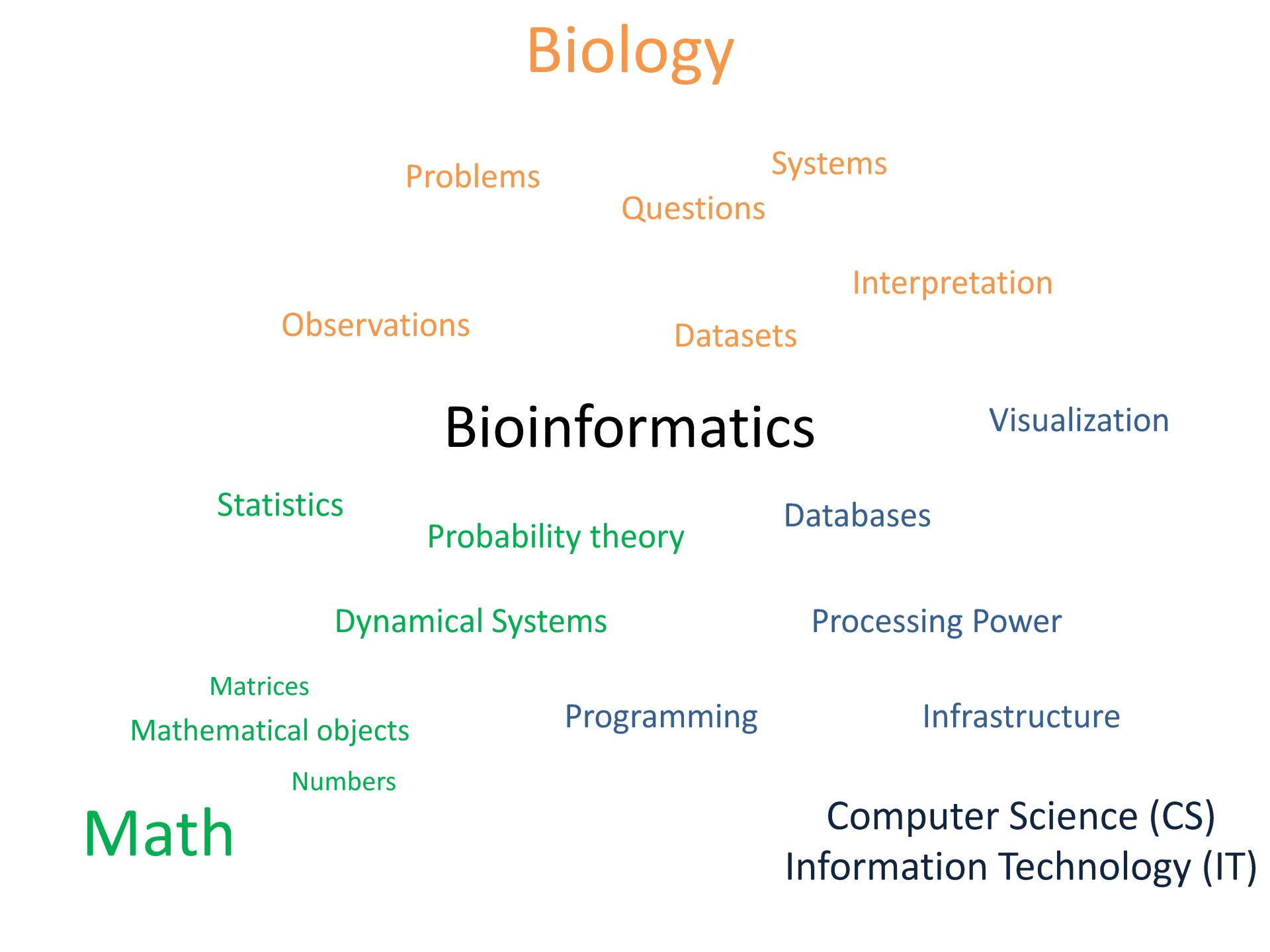
Reviews in Computational Biology

Pedro Beltrao (EMBL-EBI)

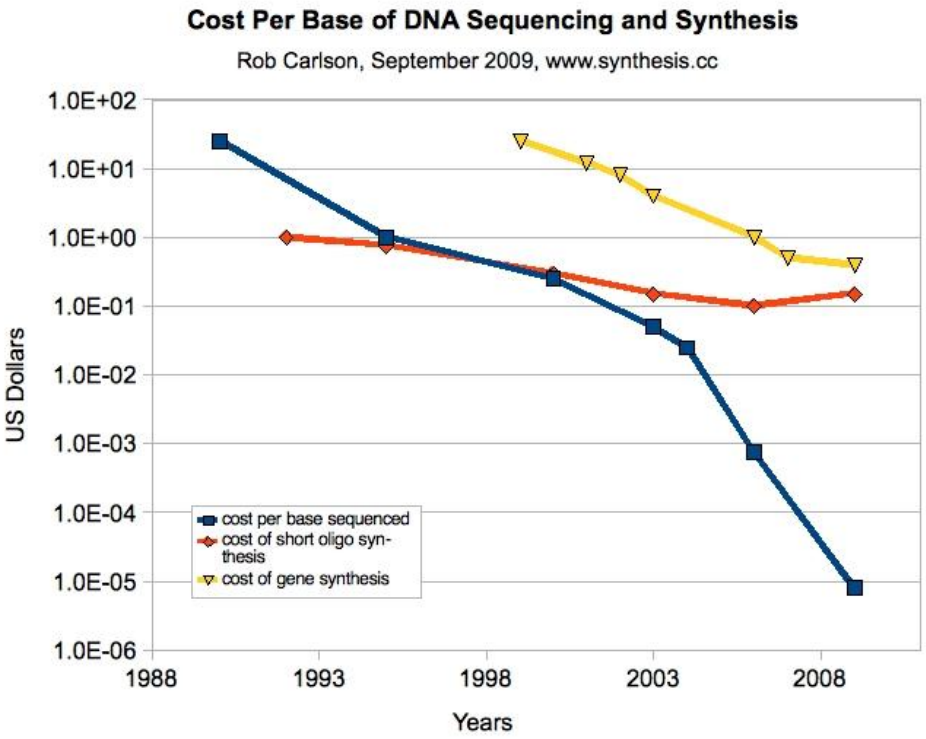
On the menu

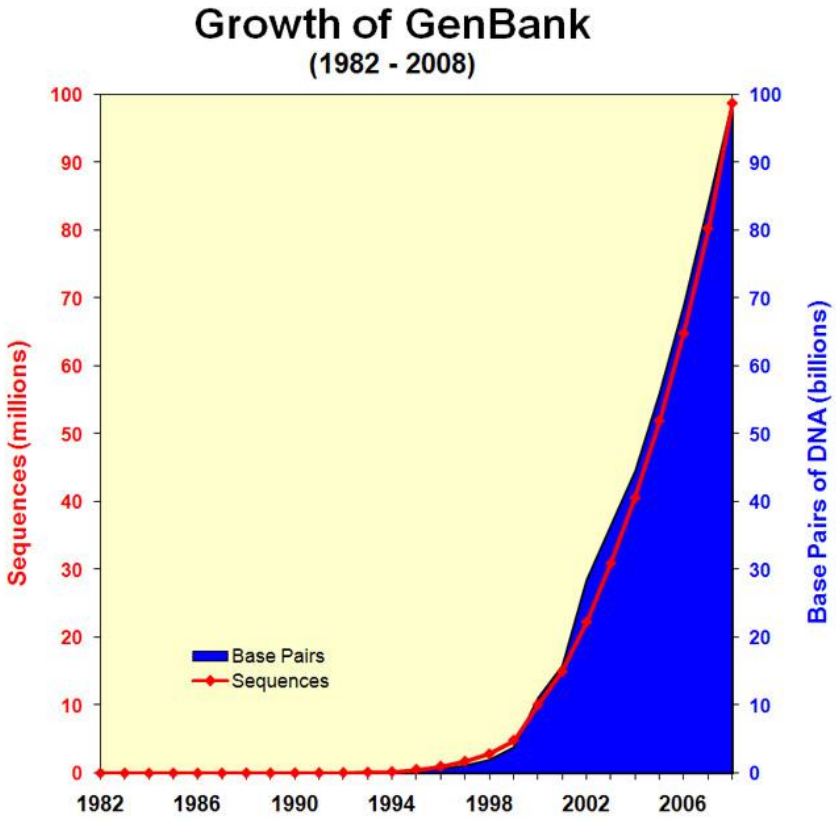
1. Very brief introduction to post-translational modifications (PTMs)
2. Evolution of PTM regulators
3. Evolution of PTM sites and interactions
4. Suggested focus and structure for the comp bio review

Bioinformatics / Comp Bio

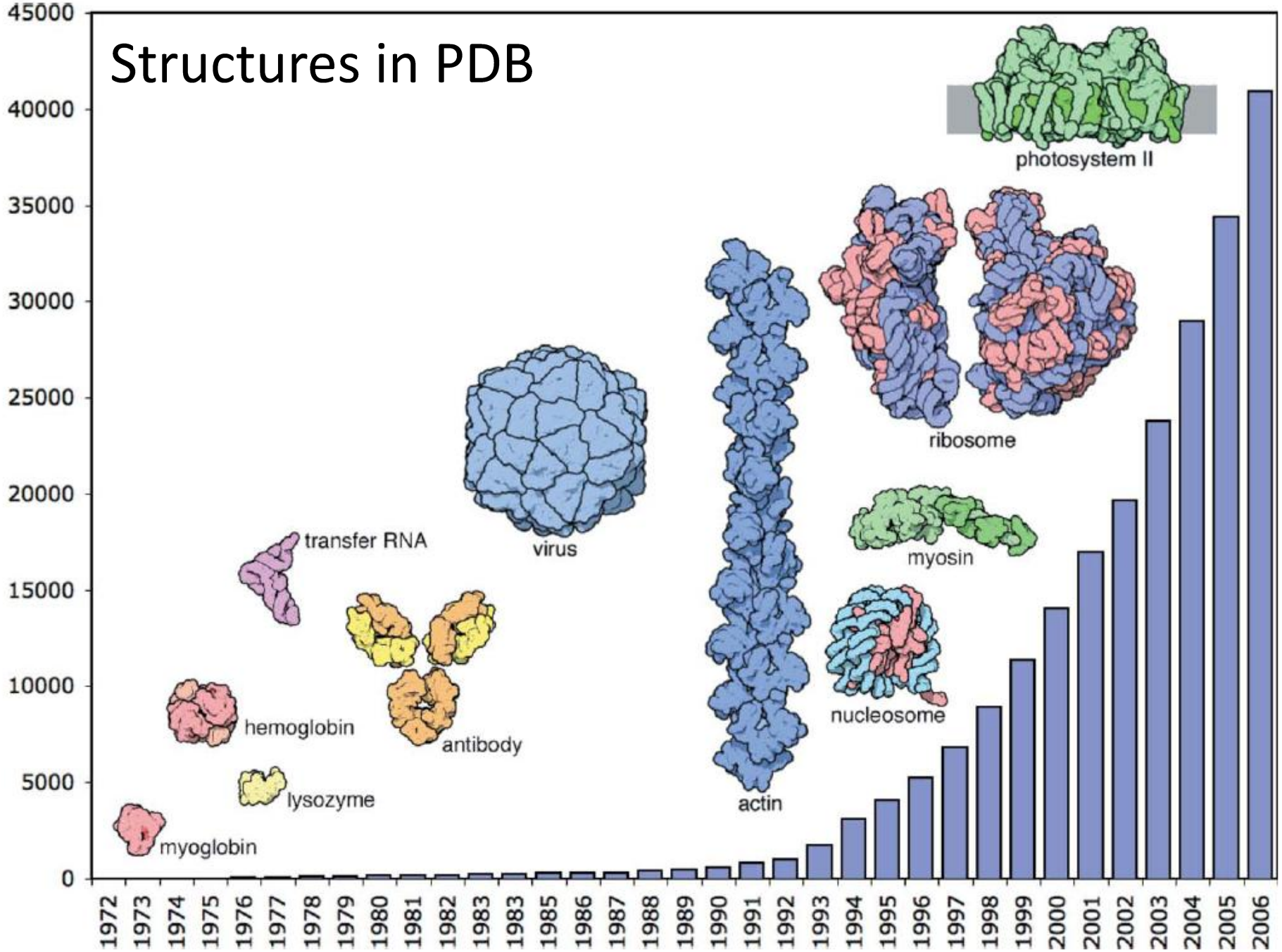

Biology
Problems
Systems
Questions
Interpretation
Observations
Datasets
Bioinformatics
Visualization
Statistics
Probability theory
Databases
Dynamical Systems
Processing Power
Matrices
Mathematical objects
Programming
Infrastructure
Numbers
Math
Computer Science (CS)
Information Technology (IT)

Exponential growth of data production

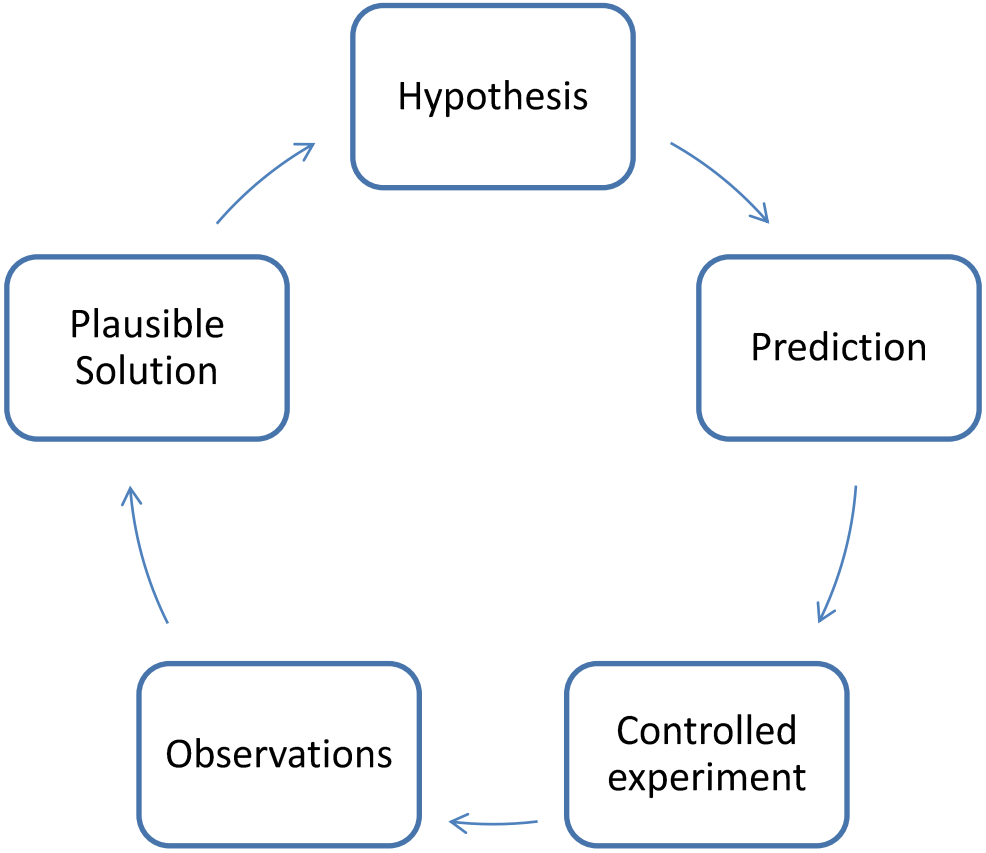




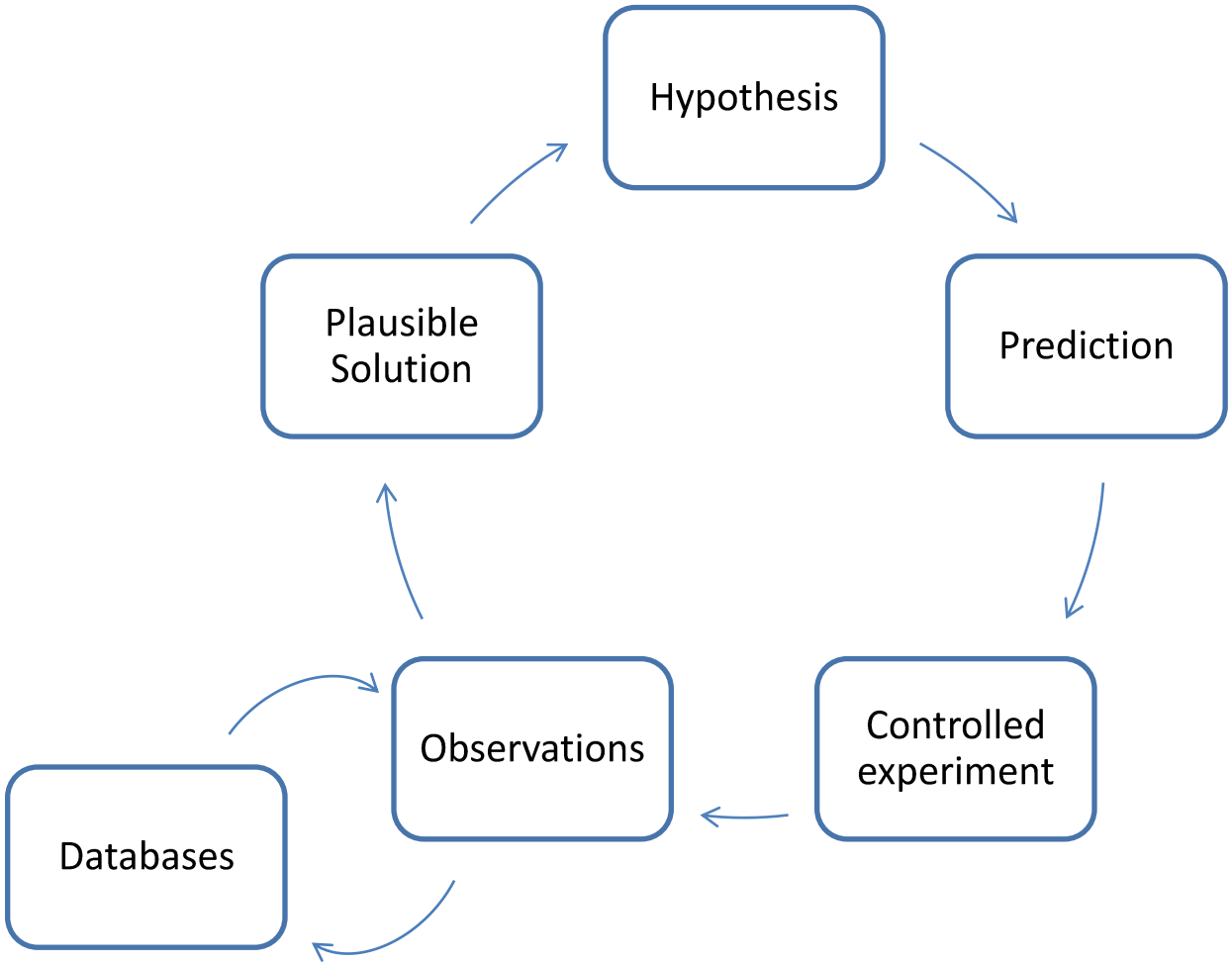
Structures in PDB



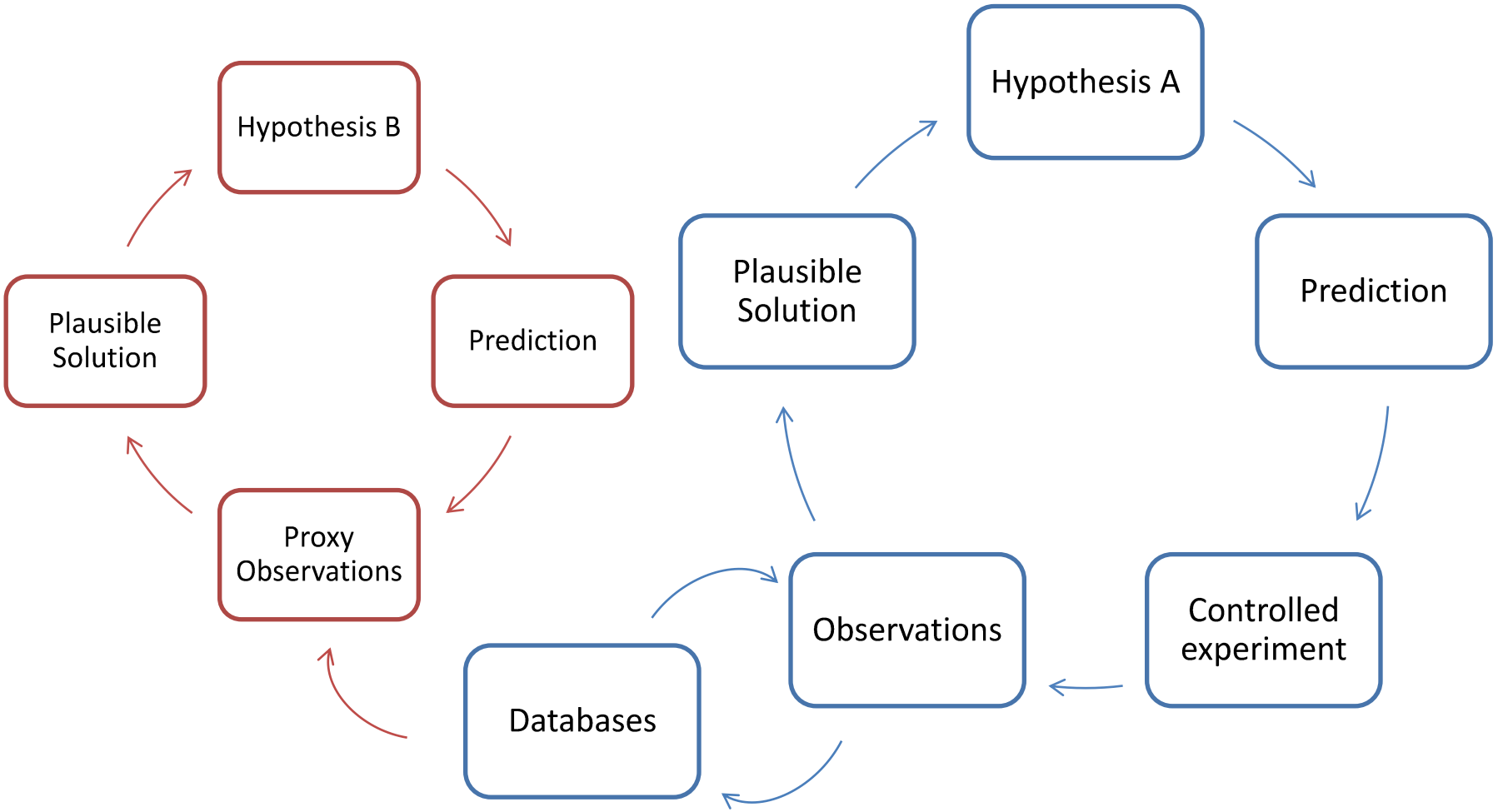
The scientific process

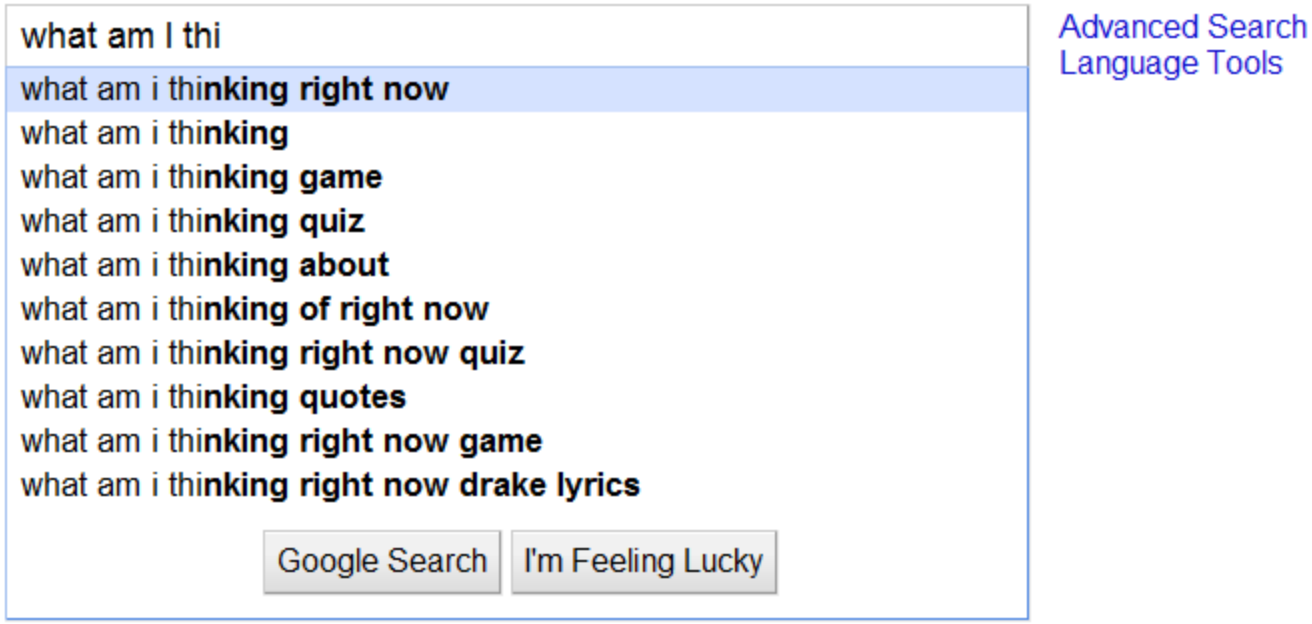


The scientific process



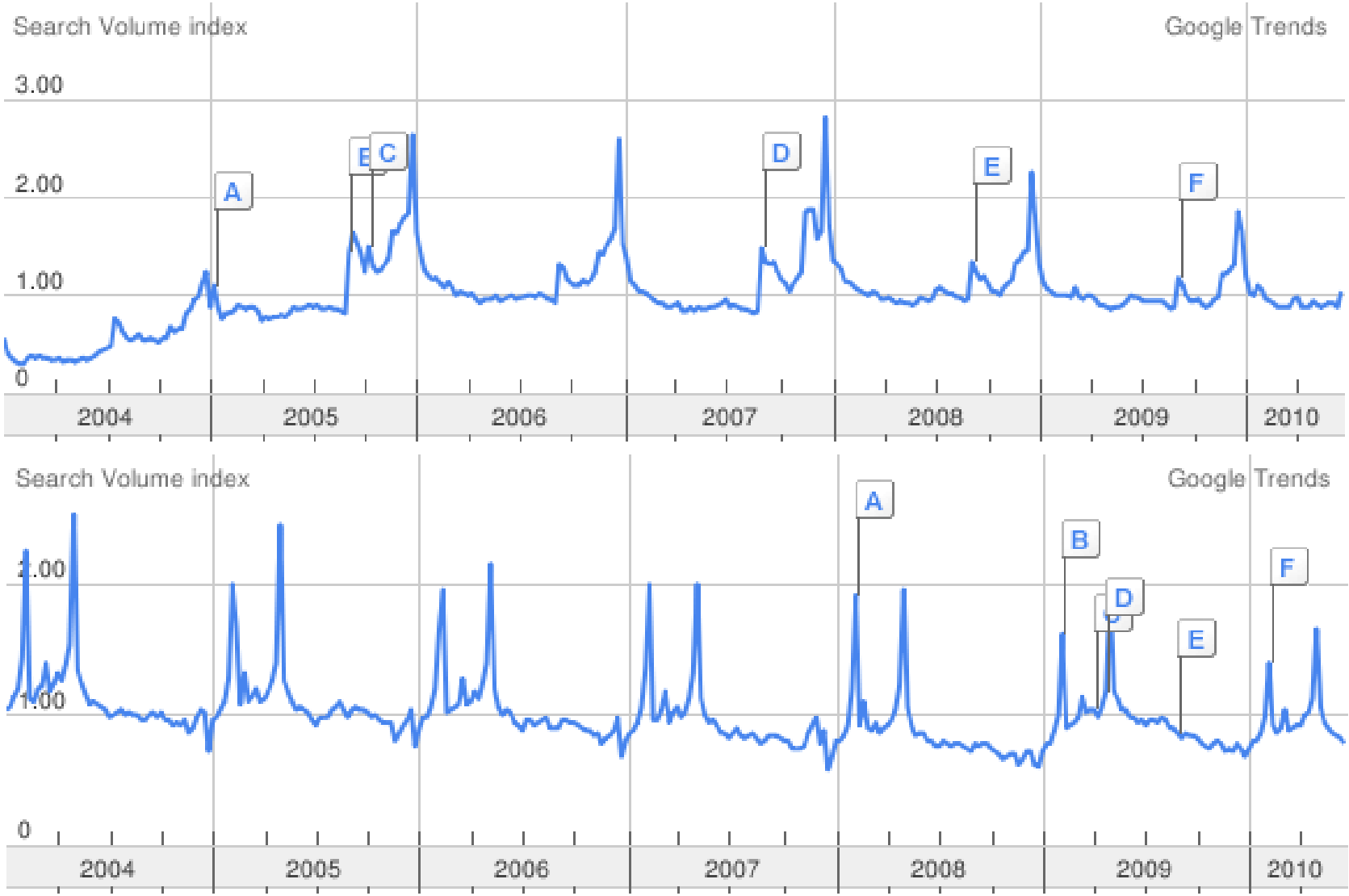
The scientific process



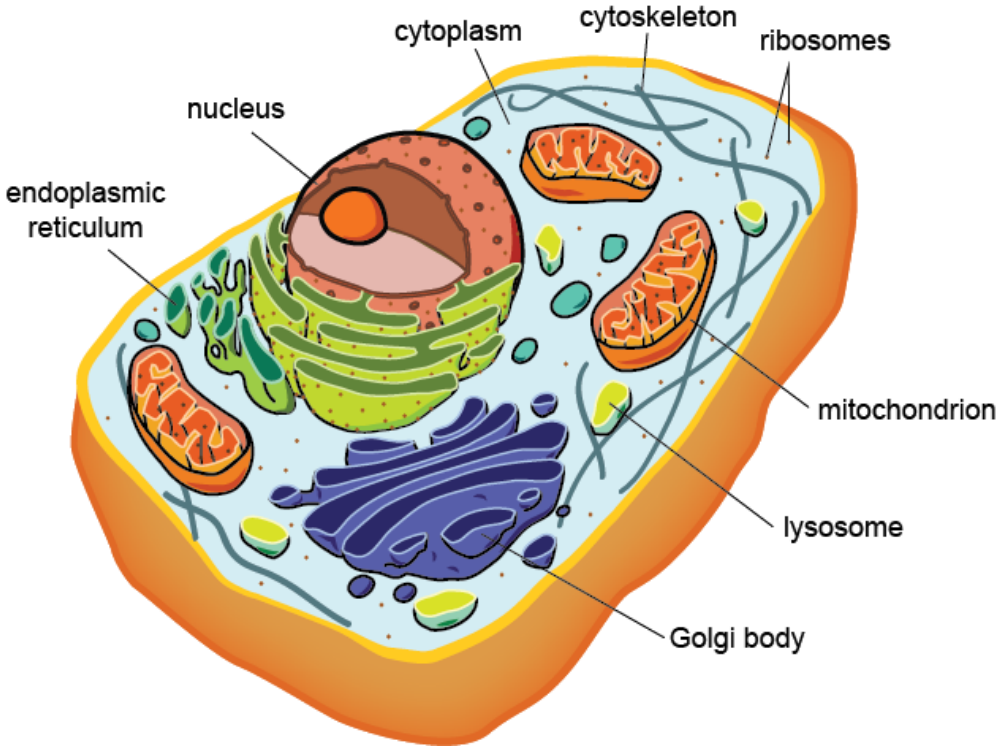


With the added benefit of having transferable skills applicable to other “big data” problems

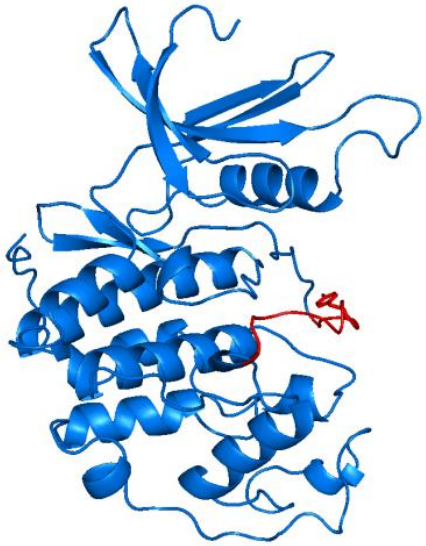
Flowers and iPods



The life of a cell



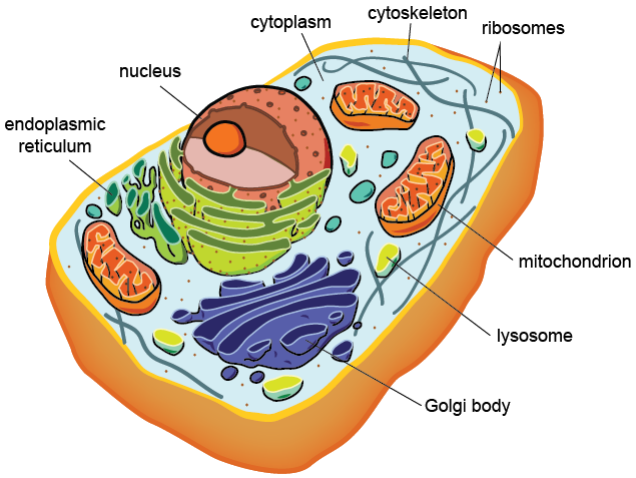
Average size of a cell is ~ 10^{-6} m



Average size of a protein is ~ 10^{-8} m

Some proteins are present with just a few copies inside the cell (transcription factors), some up to 10^{6} copies (ribosome)

The life of a cell

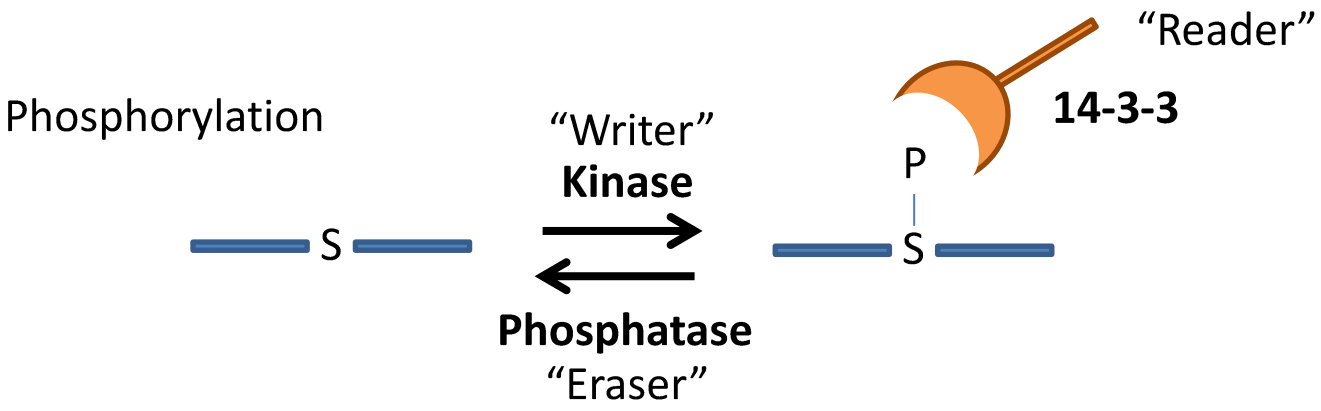


What impressions did you get about how a cell works ?

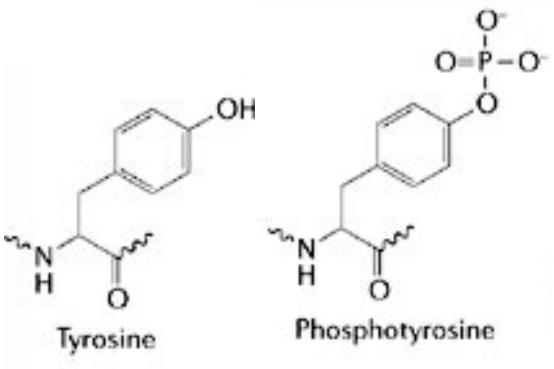
Protein Post-Translational Modifications (PTMs)

What are PTMs ?
Can you give me examples of types of PTMs ?
Why would the cell invent PTMs ?

Protein Post-Translational Modifications (PTMs)

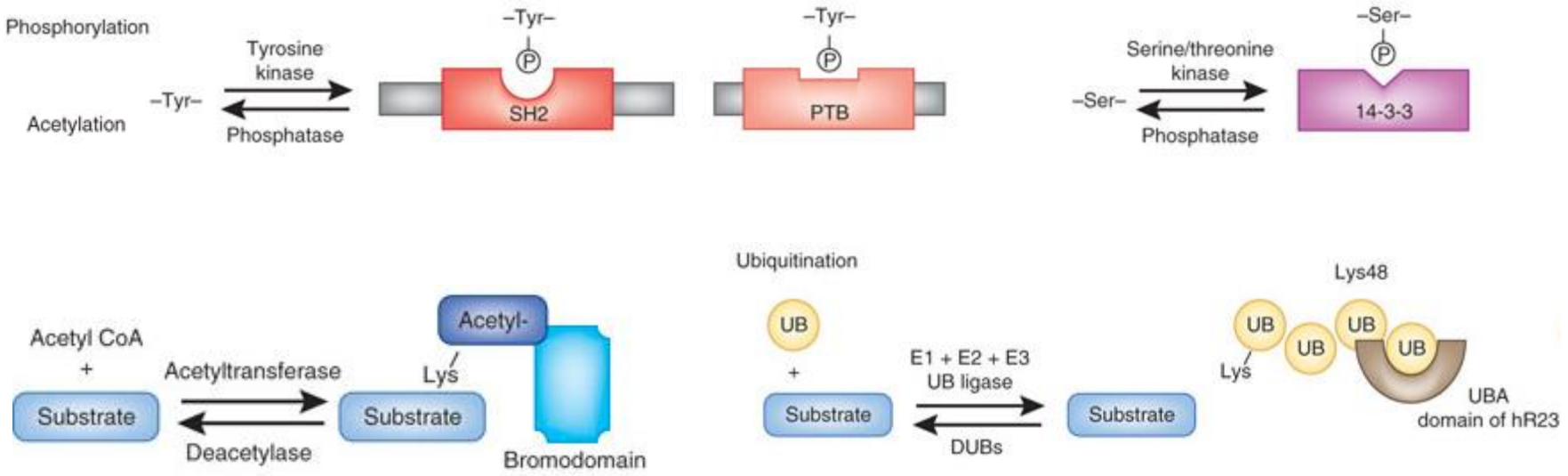


- Several amino-acids (AAs)can be phosphorylated by different types of kinases. The most commonly modified are Serine (S), Threonine (T) and Tyrosine (Y)
- The phosphorylation increases the size of the amino-acid and adds negative charge.



Reviewed in Seet et al. Nat Rev. Mol Cell Bio 2006

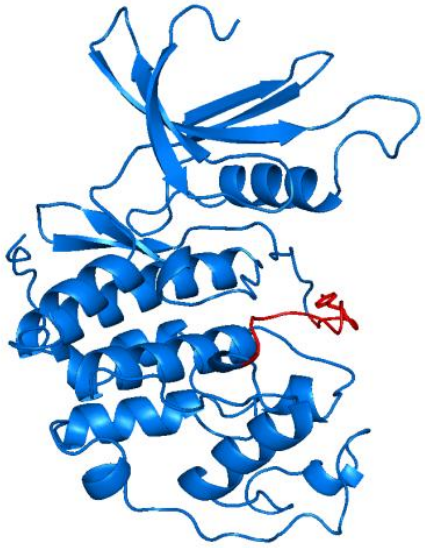
Protein Post-Translational Modifications (PTMs)



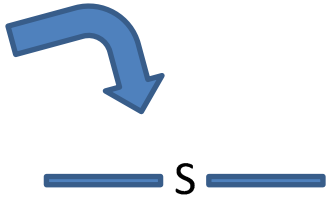
Besides regulating interactions with “reader” domains. PTMs can control function via other ways. Can you think of examples ?

Reviewed in Seet et al. Nat Rev. Mol Cell Bio 2006

Specificity of PTM regulation



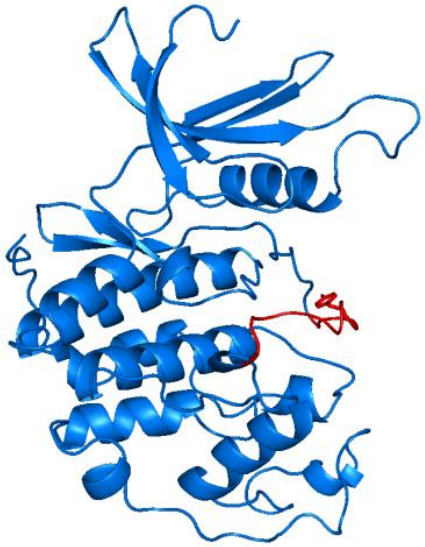
Kinase



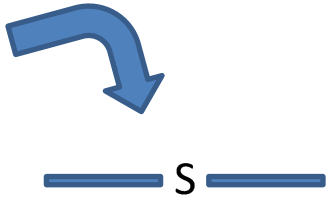
Target

There are many Serines in all the proteins inside the cell. How does a kinase know what Serines to phosphorylate ?

Specificity of PTM regulation



Kinase



Target

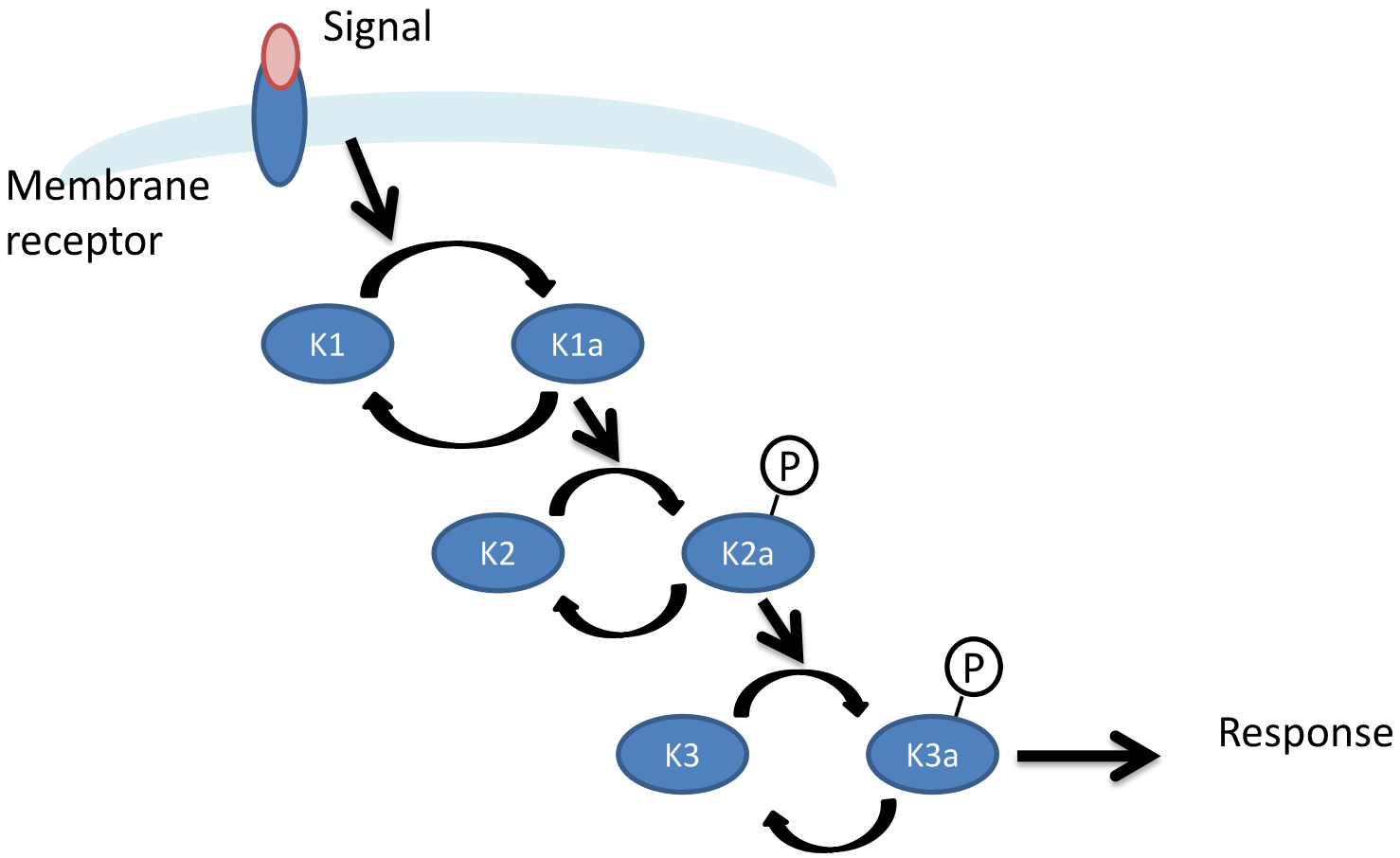
Partly explained by the presence of other amino-acids around the Serine (specificity “motifs”).

Example of kinase “motifs”

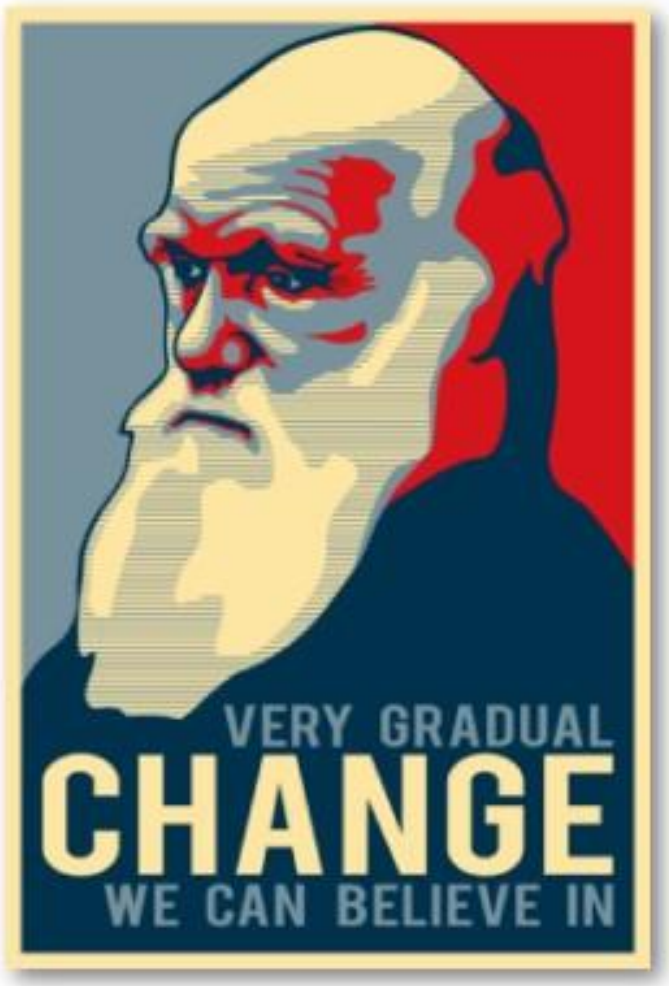
Cdk kinase - (K/R)S*PX(K/R)

These preferences are not the only specificity determinants . Other factors are important (co-localization, co-expression, interactions with common partners, etc)

Cascades of PTMs



How do species evolve ?
How are new functions created ?

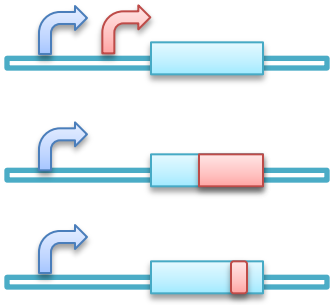


Point mutations
Recombination
Duplications

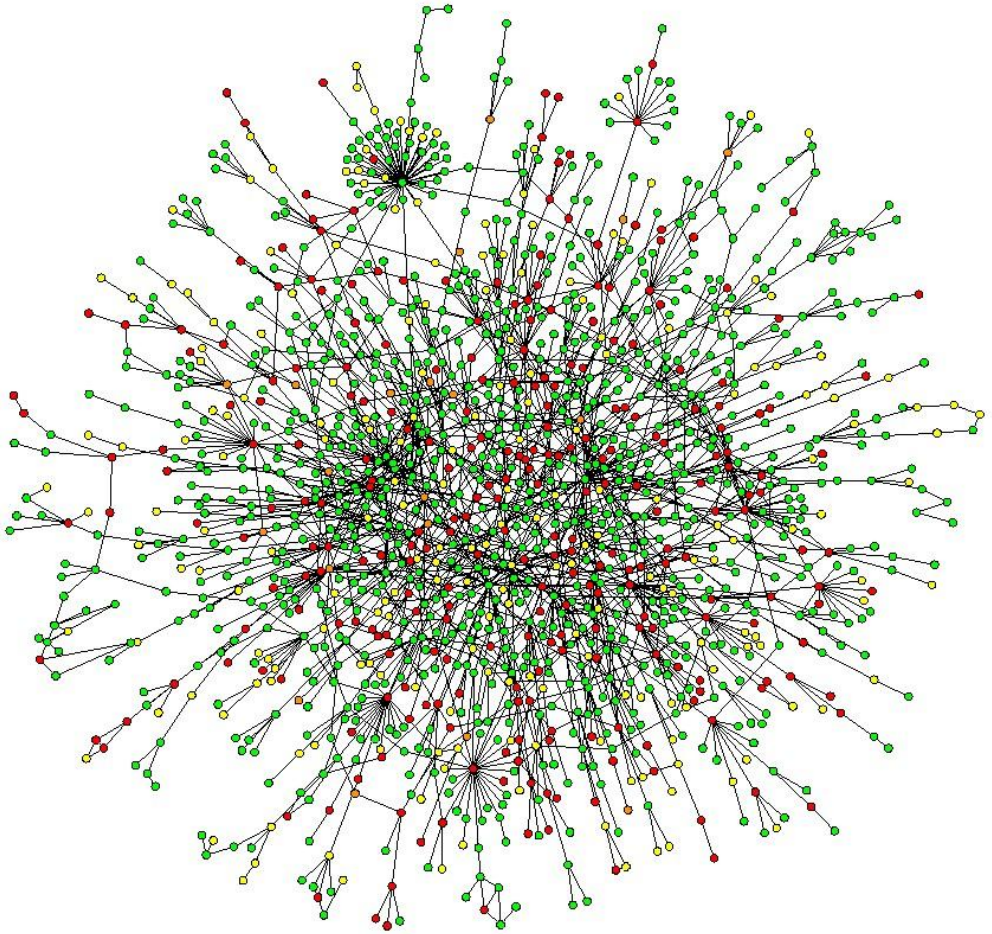
Mutants


Mutants

Point mutations
Recombination
Duplications

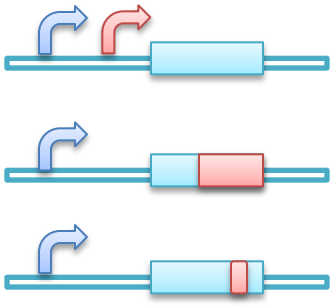


Changes in protein-protein,
protein-DNA interactions, etc

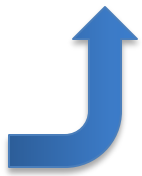


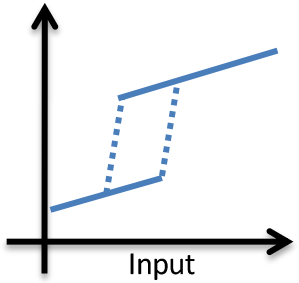
Cellular interaction networks

Point mutations
Recombination
Duplications



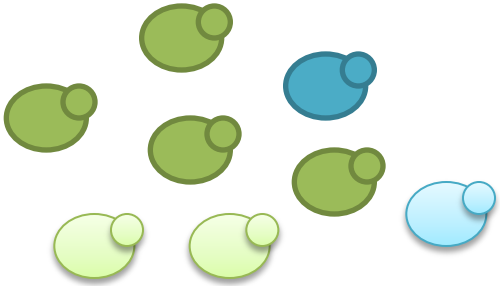
Changes in protein-protein,
protein-DNA interactions, etc



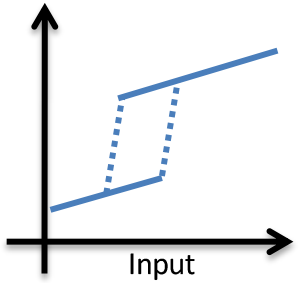


Phenotypic changes

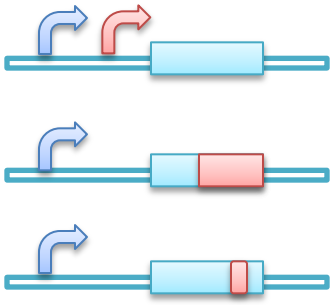
Fitness differences



Point mutations
Recombination
Duplications

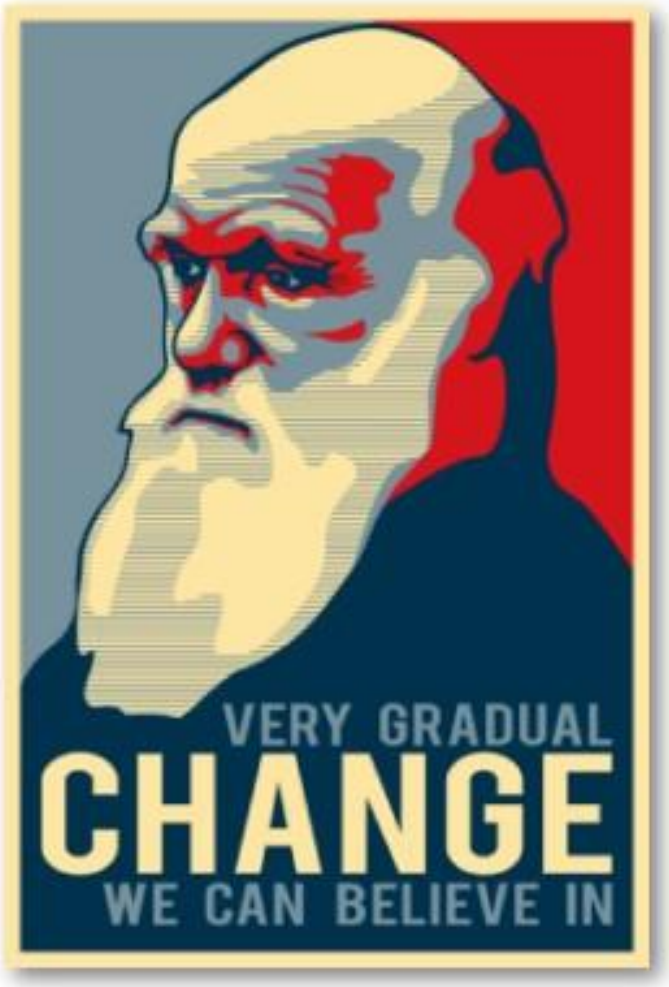


Phenotypic changes



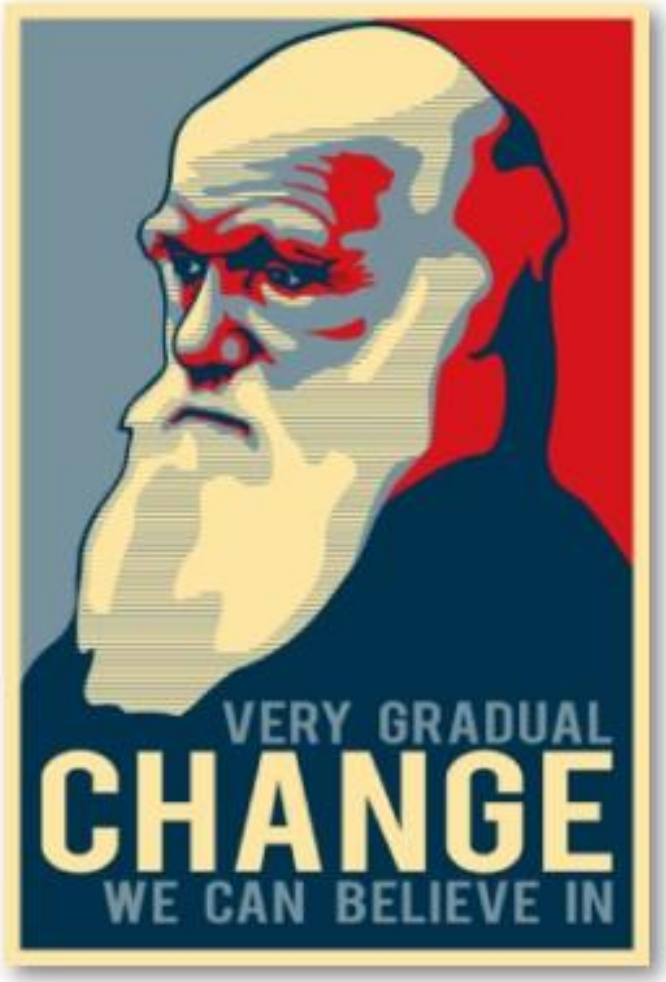
Changes in protein-protein,
protein-DNA interactions, etc

Evolution of PTMs



What questions can we ask ?

Evolution of PTMs

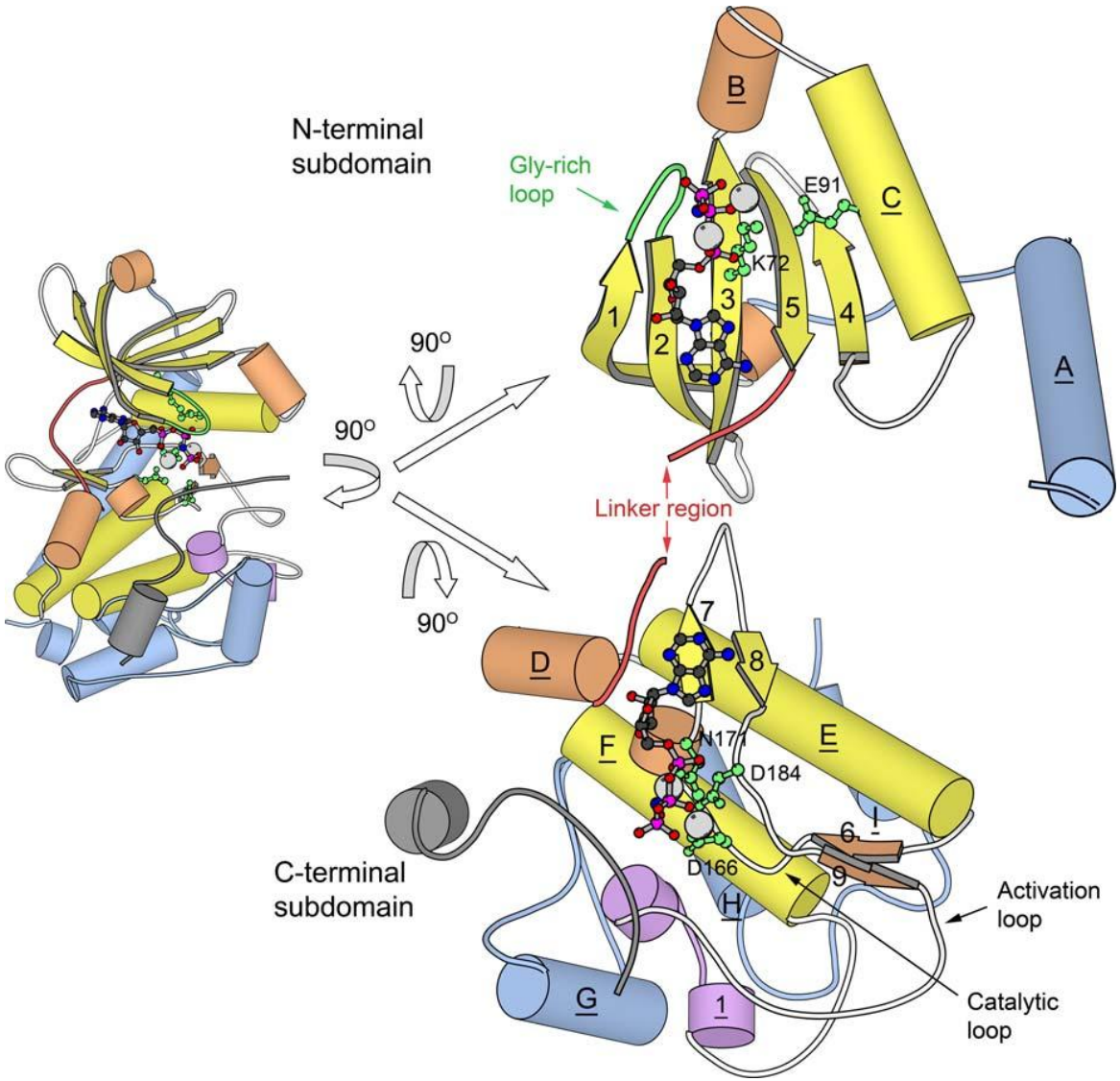


What questions can we ask ?

- What is the origin of the PTM ?
- How do the regulators change ?
- How do the targets change ?
- What changes have impact on fitness ?

I will focus on phosphorylation as the most well studied example.

Origin of Phosphorylation studied by structural homology

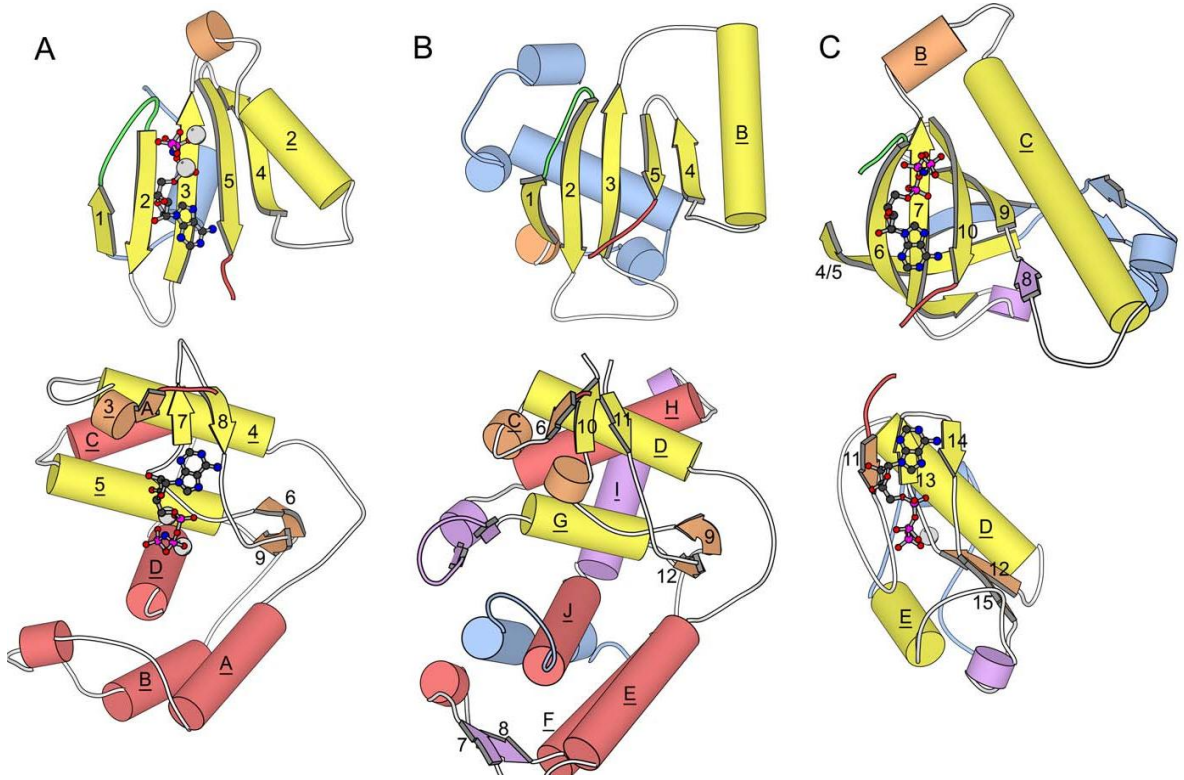


Scheeff et al PLOS Comp Bio 2005

Origin of Phosphorylation studied by structural homology

N-terminal subdomain

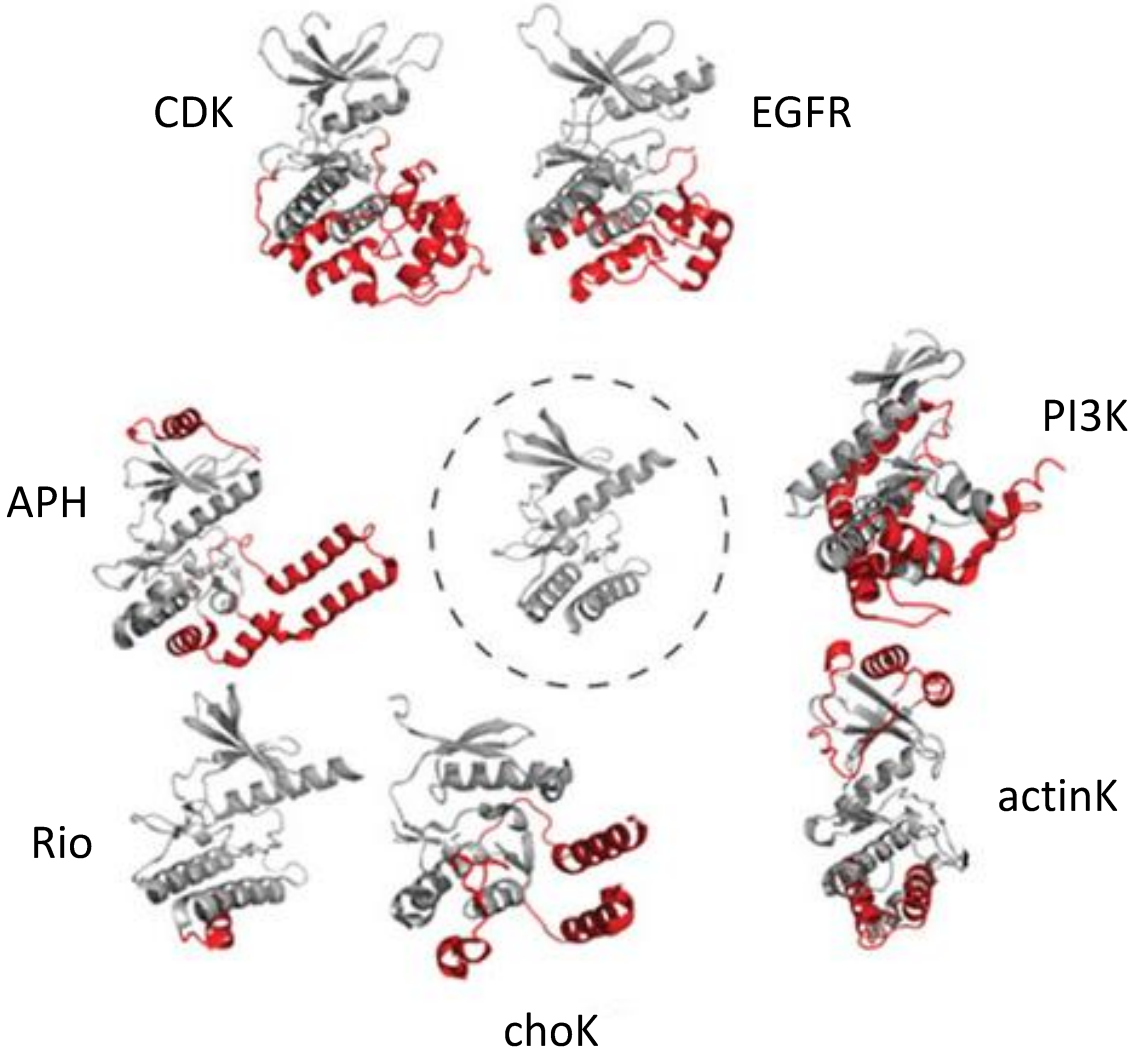
C-terminal subdomain



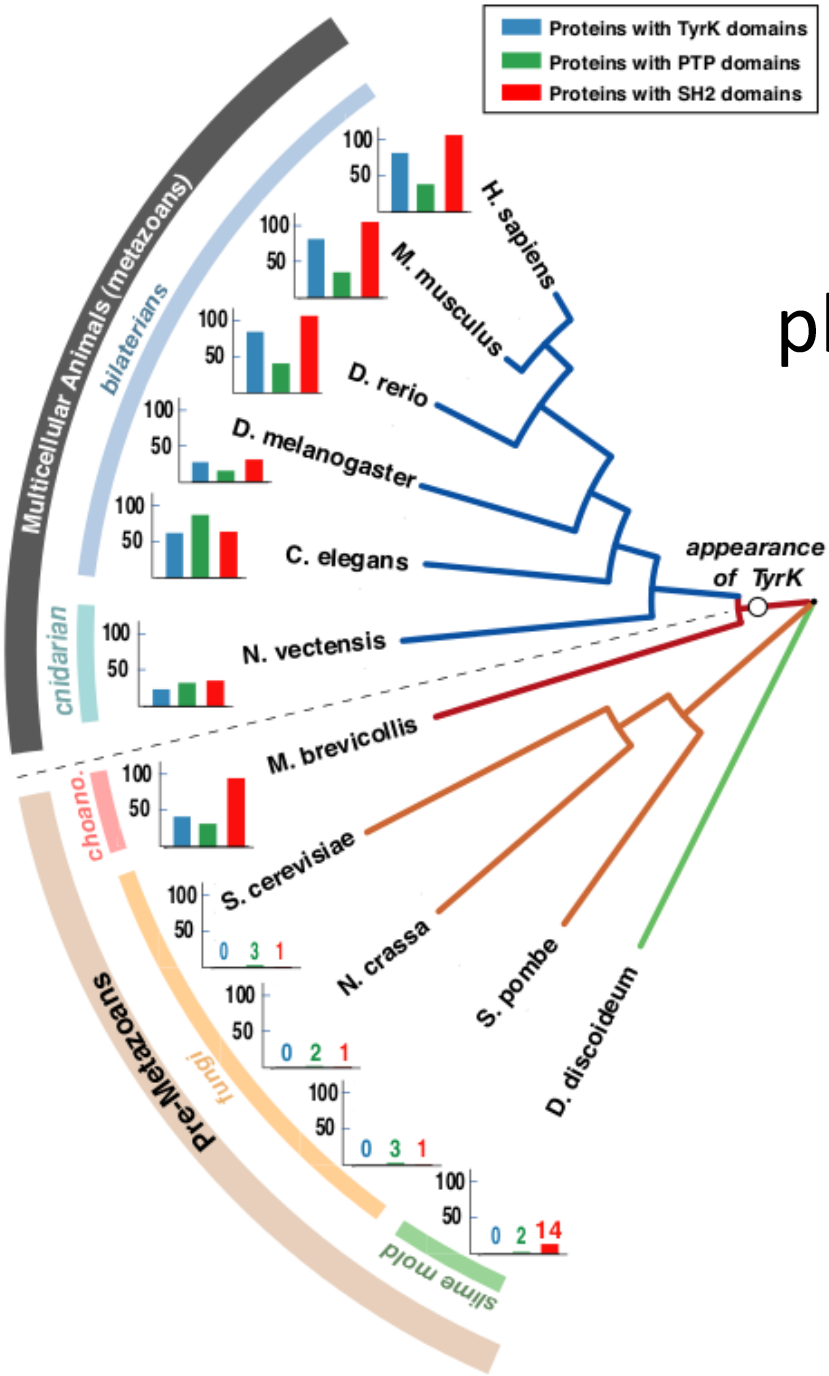
Protein kinases exist across all life but have been expanded significantly in eukaryotic species. There are kinases that phosphorylate small-molecules and lipids. It is hard to study the origin of these very ancient events

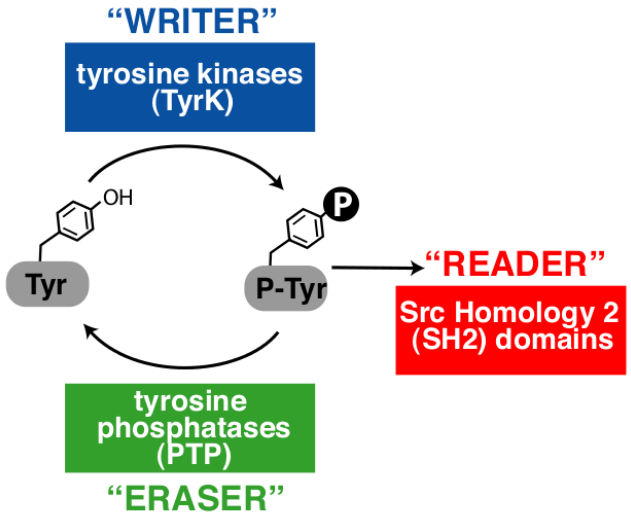
Scheeff et al PLOS Comp Bio 2005

Diversification of protein kinases (through gene duplication and divergence)



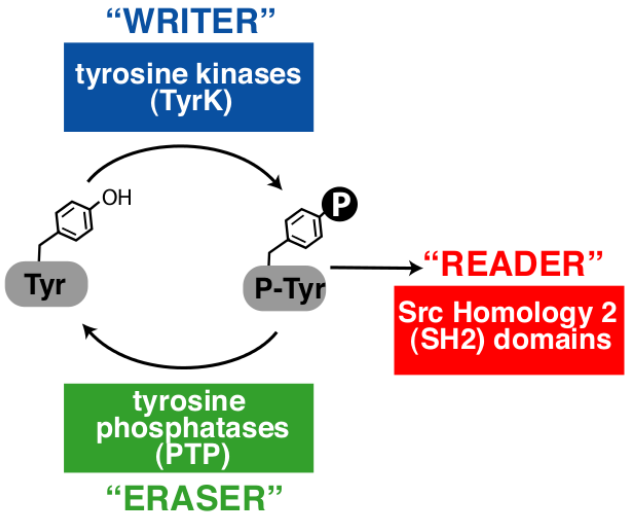
Evolution of the tyrosine phosphorylation machinery





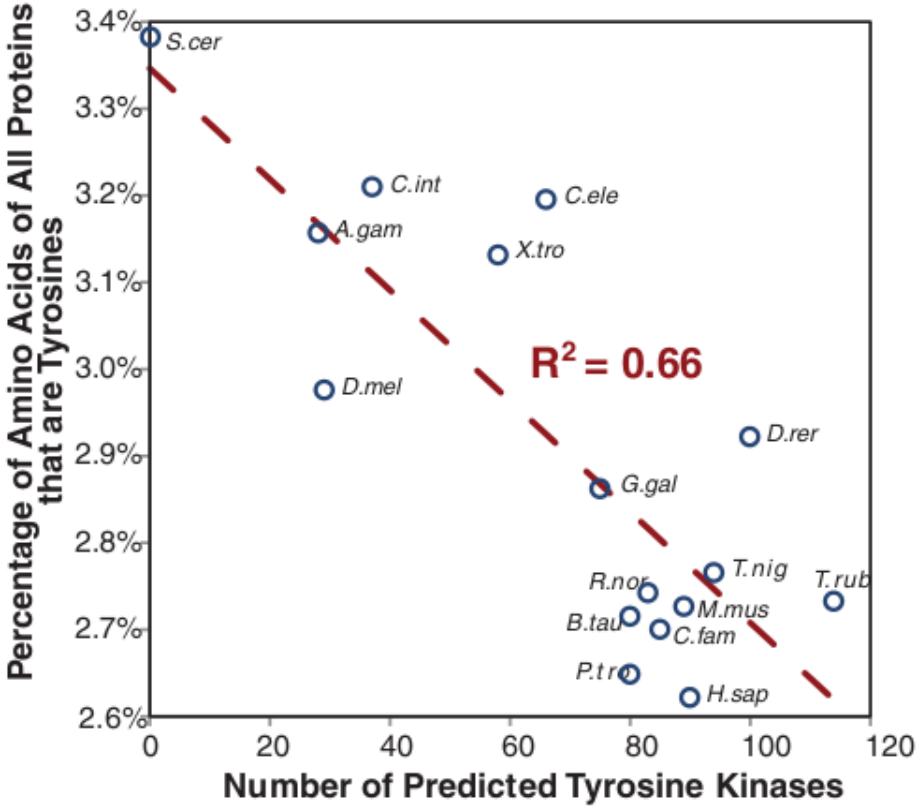
Pincus et al. PNAS 2008

Costs of having a new modification



What sort of problems did the first tyrosine kinases create after they were accidently invented by cells ?

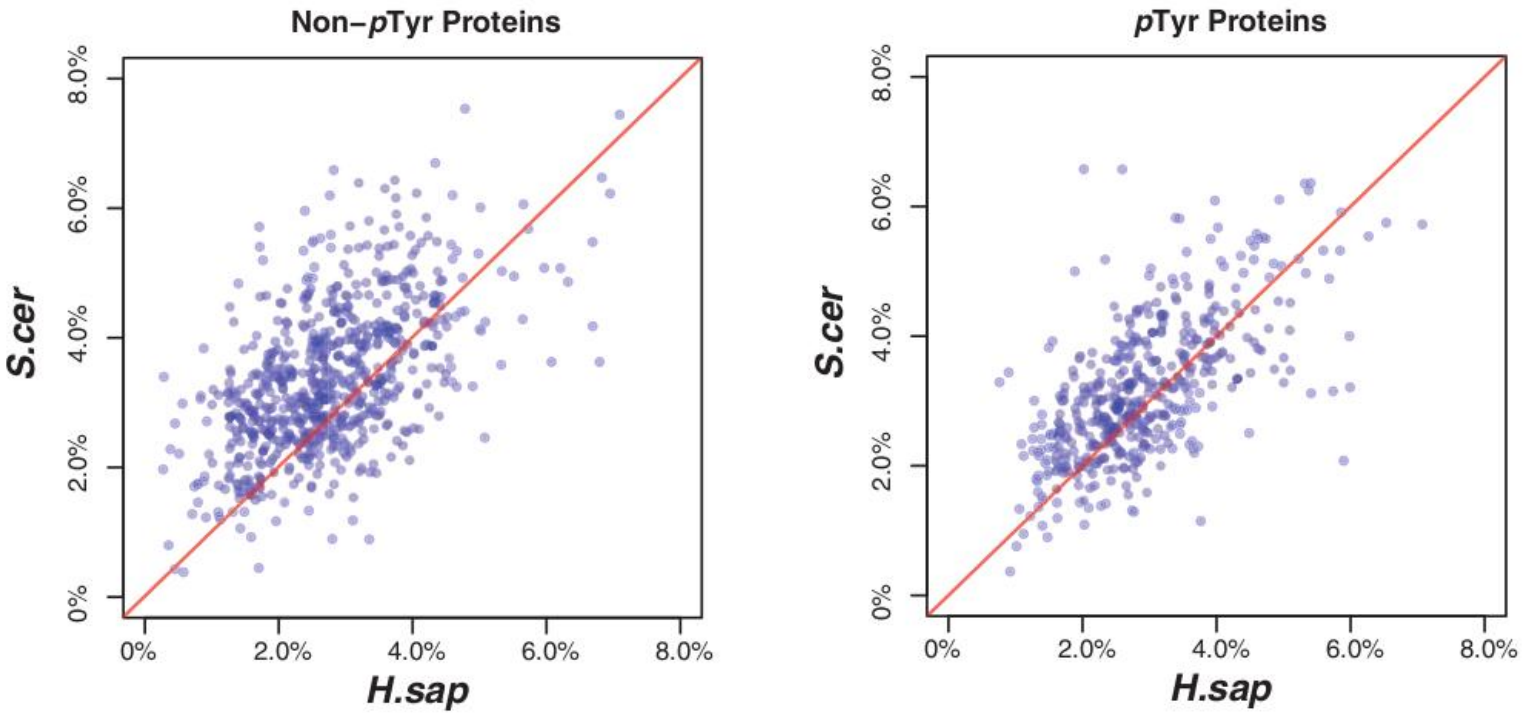
Costs of having a new modification



Species that have more tyrosine kinases have proteins with lower percentage of tyrosines

Tan et al. Science 2009

Costs of having a new modification



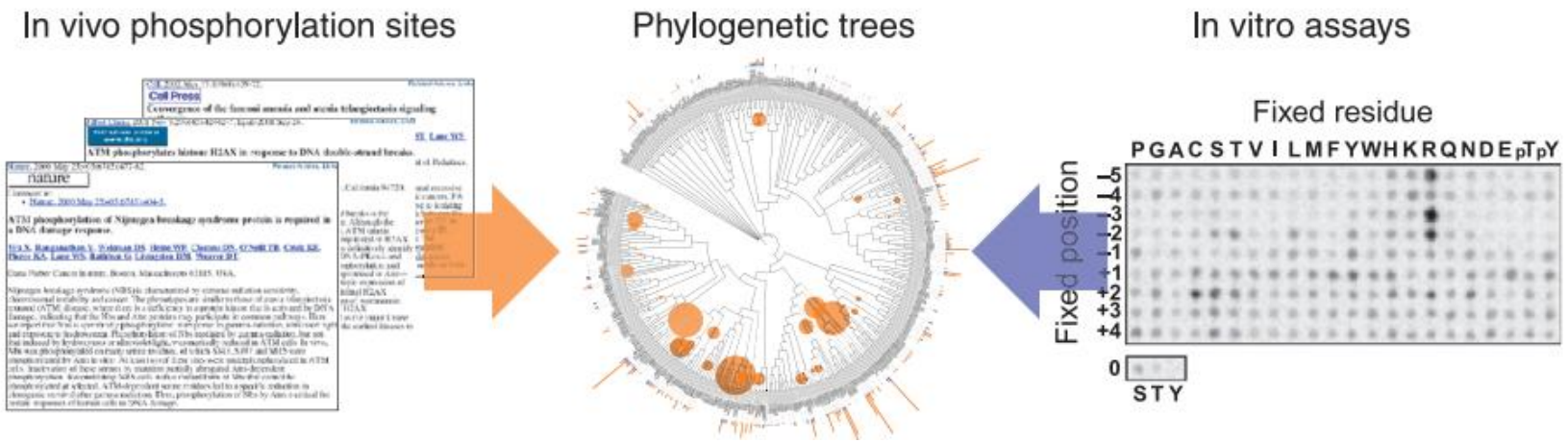
The reduction of tyrosines (in the yeast orthologs) is more noticeable in proteins that are today phospho-tyrosine proteins in human.
Suggests: Some tyrosine phosphorylation was detrimental and there was selective pressure to mutate these to other amino-acids.

Tan et al. Science 2009

Predicting the specificity of the subfamilies

- Each kinase sub-family and (presumably) each kinase protein has a preference for specific targets. The specificity of the binding site can be predicted by two different approaches:
  - Machine learning applied to set of known target sites (e.g. Netphorest)
  - Structural analysis of the binding interfaces (e.g. Predikin)

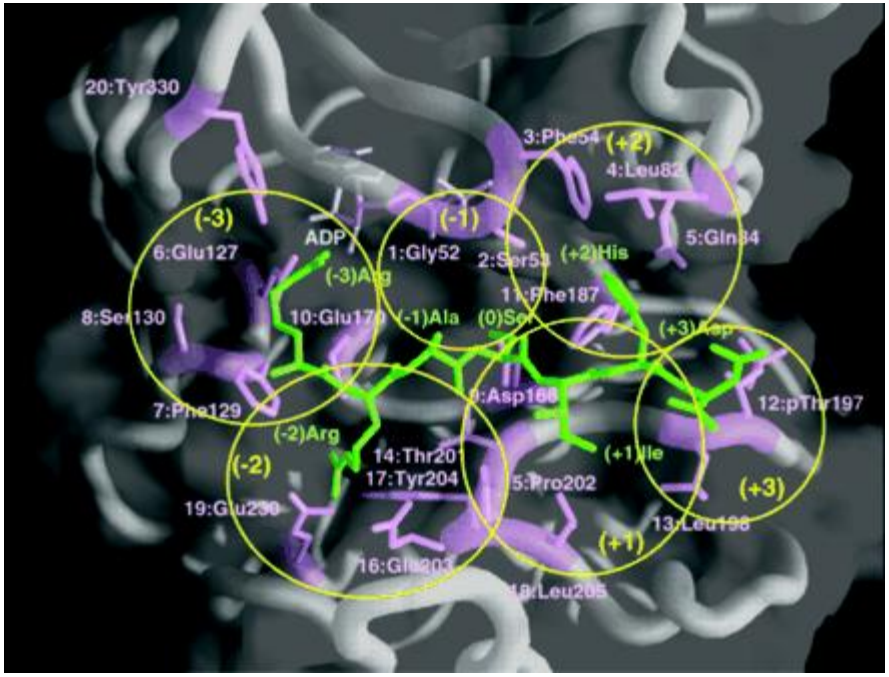
Machine learning applied to set of known target sites (e.g. Netphorest)



After reducing potential redundancies these were used as Training/test sets for data mining

Miller et al. Science Signaling 2008

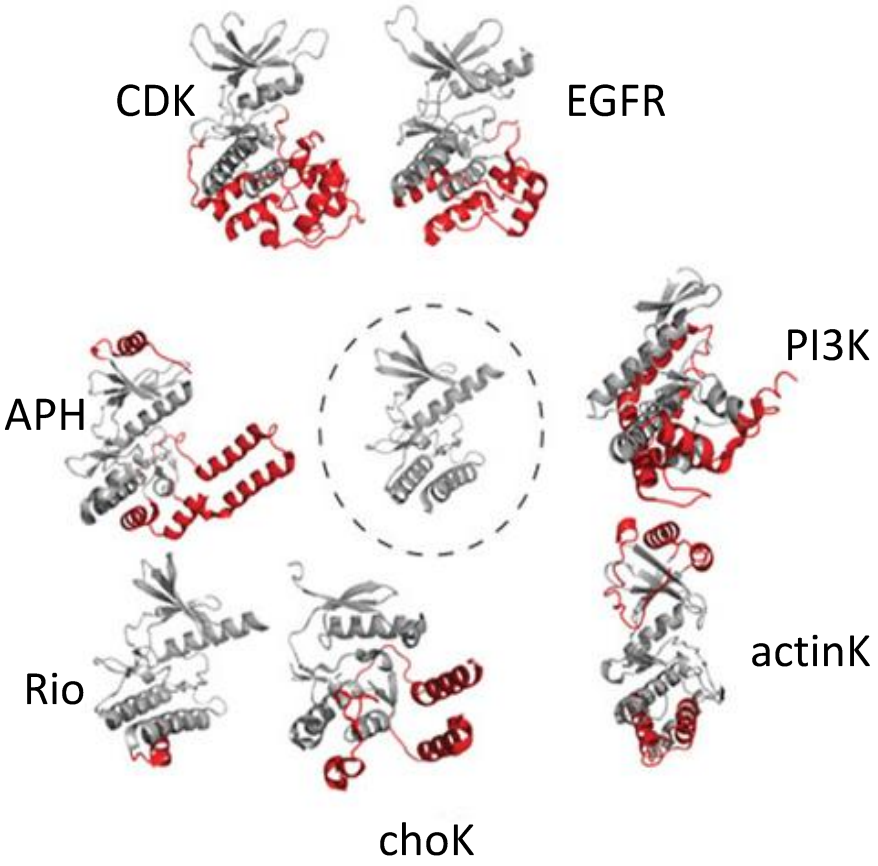
Structural based prediction (e.g. Predikin)



- Compiled a set of structures of kinases bound to substrate sites.
- Extracted the residues that were in contact with the peptide.
- Correlated changes in kinase residues with preferences for the target “motif”
- Defined a set of rules that can be applied to different kinases based on sequence

Brinkworth et al. PNAS 2002
Ellis et al. PLOS One 2011

Diversification of protein kinases (through gene duplication and divergence)



- These methods have so far not been applied to study the evolution of binding specificity
- They have been used to predict interactions between kinases and phosphosites and to study their evolution (in next section)

Recap – Evolution of regulators

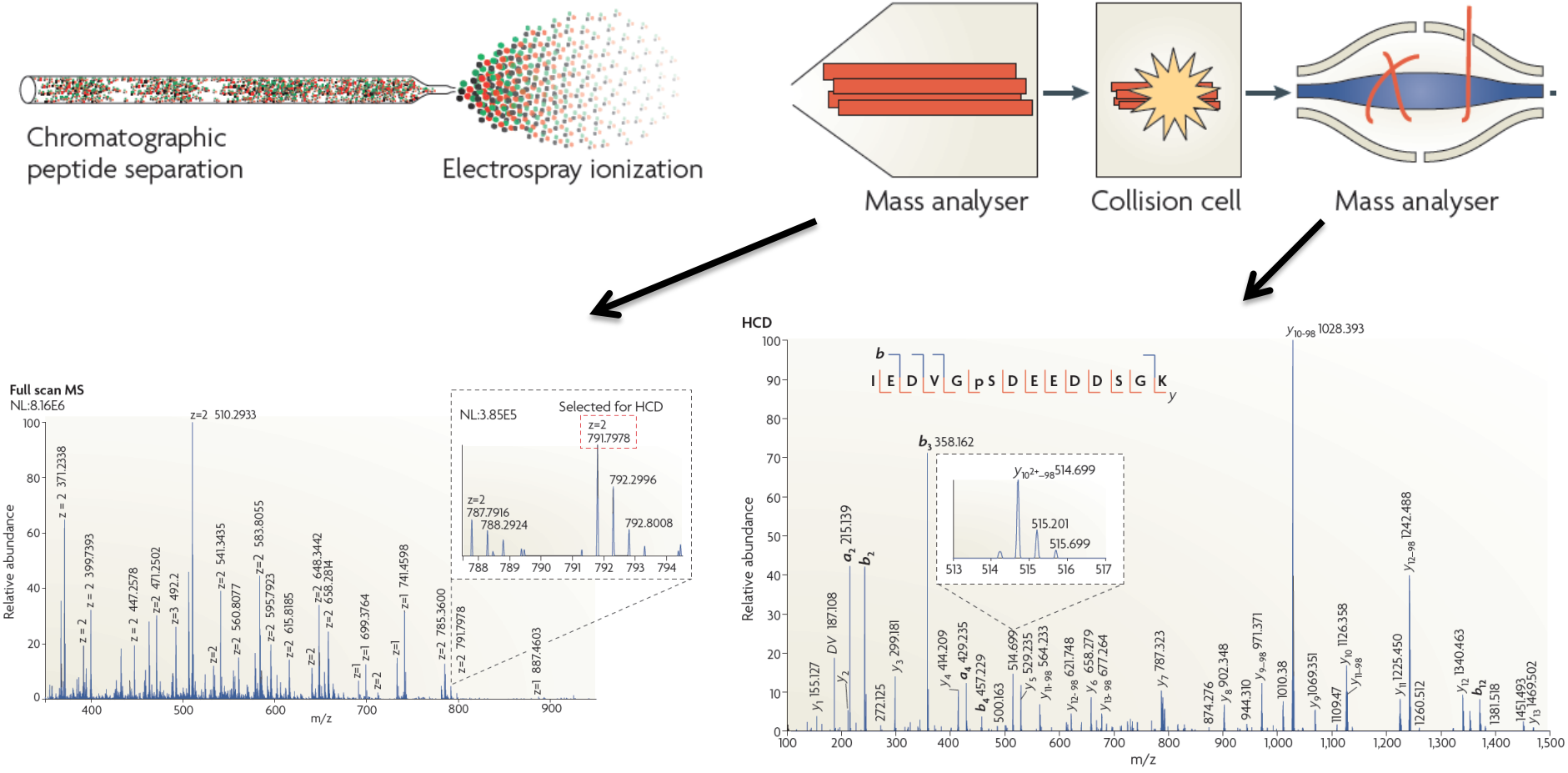
- Novel PTMs are created very rarely and their ancient origin is hard to study
- The usefulness of the PTM depends to some extent of having a set of 3 different types of proteins (writers, erasers and readers).
  - The origin of tyrosine phosphorylation appears to have coincided with a burst of duplication of the 3 types of domains
- Novel PTMs are likely to introduce pressures/costs since the proteome is not ready to have a large amount of the novel modification
- The PTM enzymes, once “invented” diverge by duplication and divergence into sub-families that have their own specificities
  - Kinase sub-families prefer to phosphorylate different amino-acid “motifs” around the phosphosite (ex. RS*PXK)
  - Kinases do not appear to be very specific and other types of information (co-localization, co-expression, etc) are important for recognition

Recap – Evolution of regulators (Computational Methods)

- The study of the very ancient evolutionary origins of PTMs depends on the analysis of structural homology.
- The duplication/divergence into subfamilies is commonly studied by training domain family models (ex PFAM database). These are used to search across multiple species.
- Structural bioinformatics or machine learning methods can be used to predict the specificity of kinases

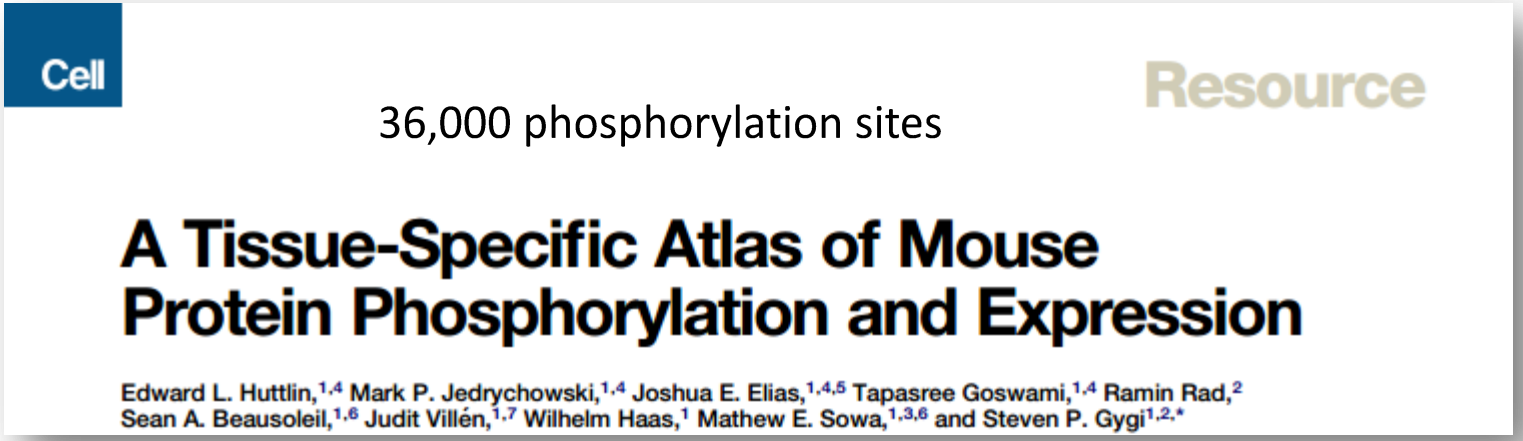
Evolution of PTM sites and interactions

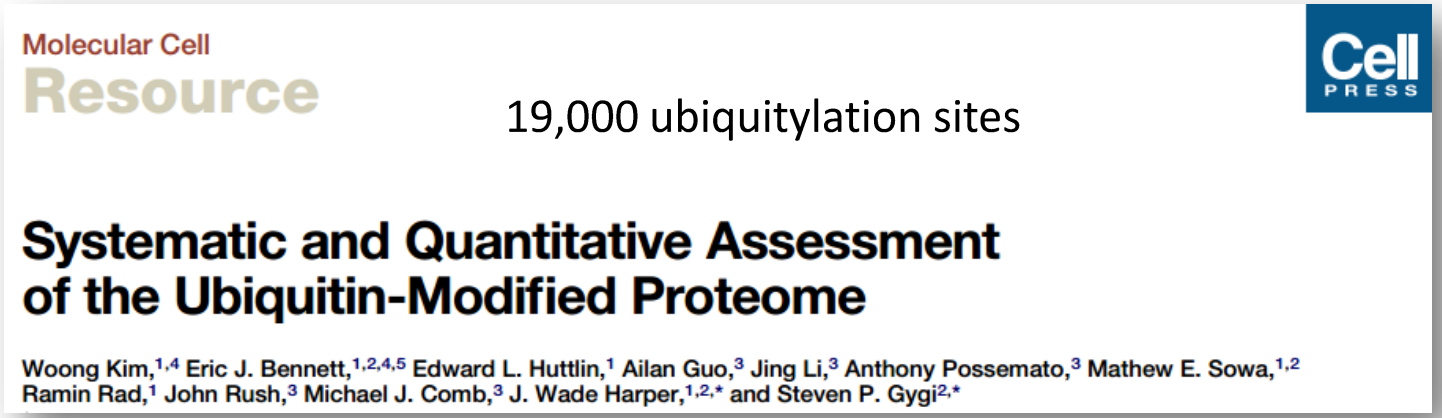
PTMs can be experimentally identified using mass-spectrometry



PTMs can be experimentally identified using mass-spectrometry

1. Sample preparation
    1. Extraction
    2. Purification
    3. Protein digestion (**peptides**)
    4. Enrichment for PTM (antibody or binding column)
2. Tandem Mass-spec (MS/MS)
    1. Chromatographic peptide separation
    2. Ionization
    3. Peptide ions mass/charge (m/z) determination
    4. Peptide fragmentation
    5. Fragment ions m/z determination
3. Computational analysis
    1. Compare experimental spectra with predicted spectra
    2. Determine false discovery rate

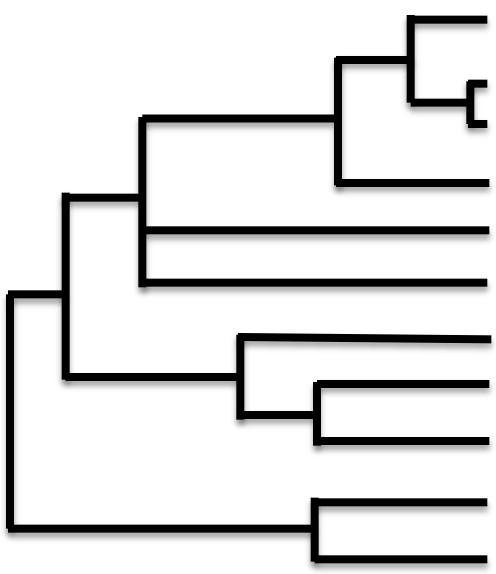




The enrichment methods and sensitivity of the MS machines has improved tremendously in the past 5 years.

Known eukaryotic PTM sites

(curated from MS studies)

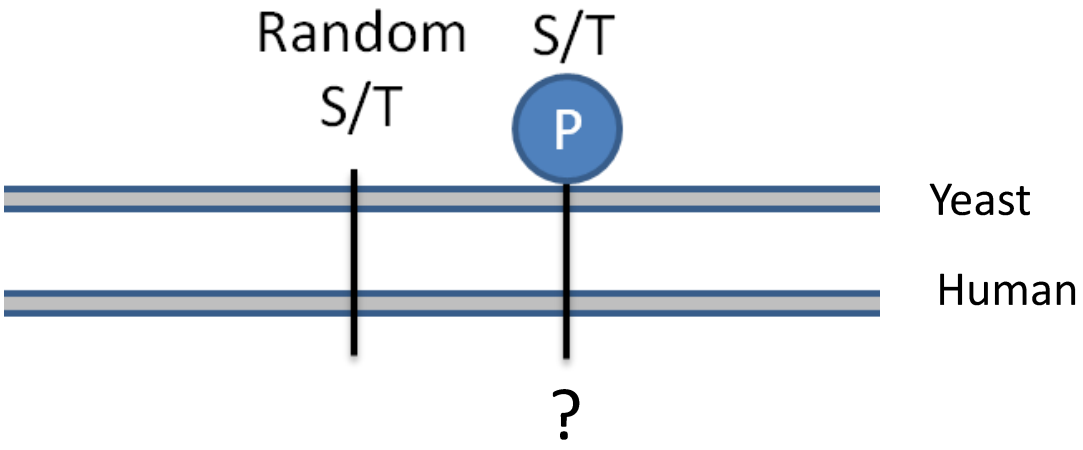


Species	Phospho sites	Acetylation	Ubiquitylation
H.sapiens	45943	8042	22000
M.musculus	32885	3384	
R. norvegicus	2702		
X.laevis	629		
C.elegans	8433		
D.melanogaster	22307	1707	
S.pombe	2989		
S.cerevisiae	22536		5000
C.albicans	3773		
A.thaliana	5336		
O.sativa	3666		

> 150.000 phosphorylation sites

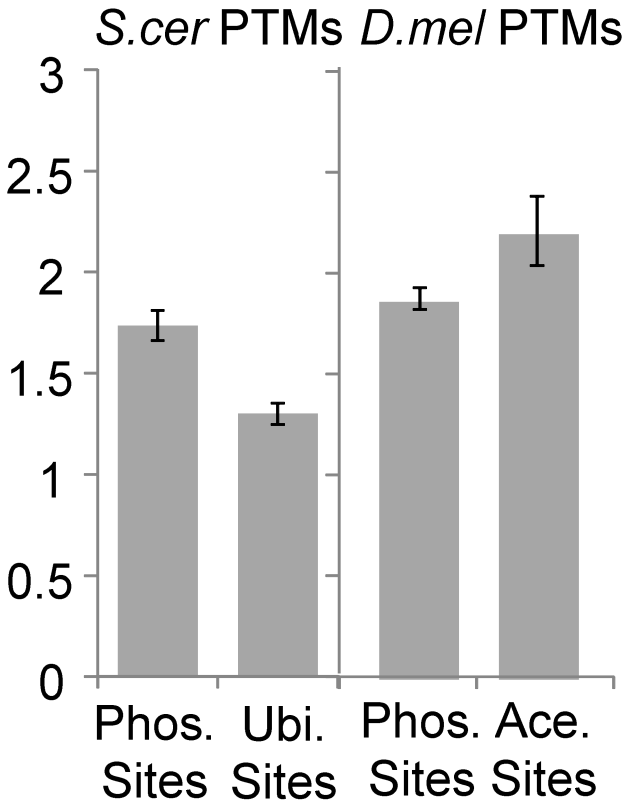
~ 200.000 PTMs

Measuring conservation



- Conservation of “State” (ex. phosphorylated): The same peptide in the human protein is also phosphorylated
- Compared with random sampling of acceptor residues in the same proteins - ratio of conservation over random.

Conservation of post-translational modifications



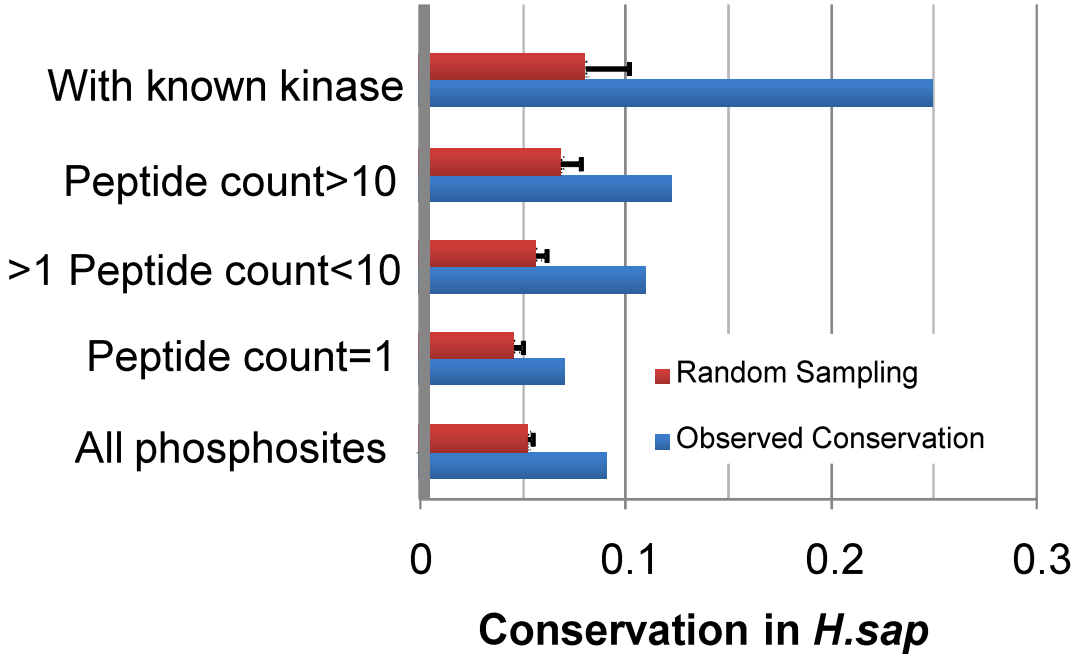
All 3 modifications are poorly evolutionarily constrained when compared to random sample of acceptor residues

If there is a difference:
AceK > phospho > ubi

Suggest that ubiquitylation is less specific than the other 2.

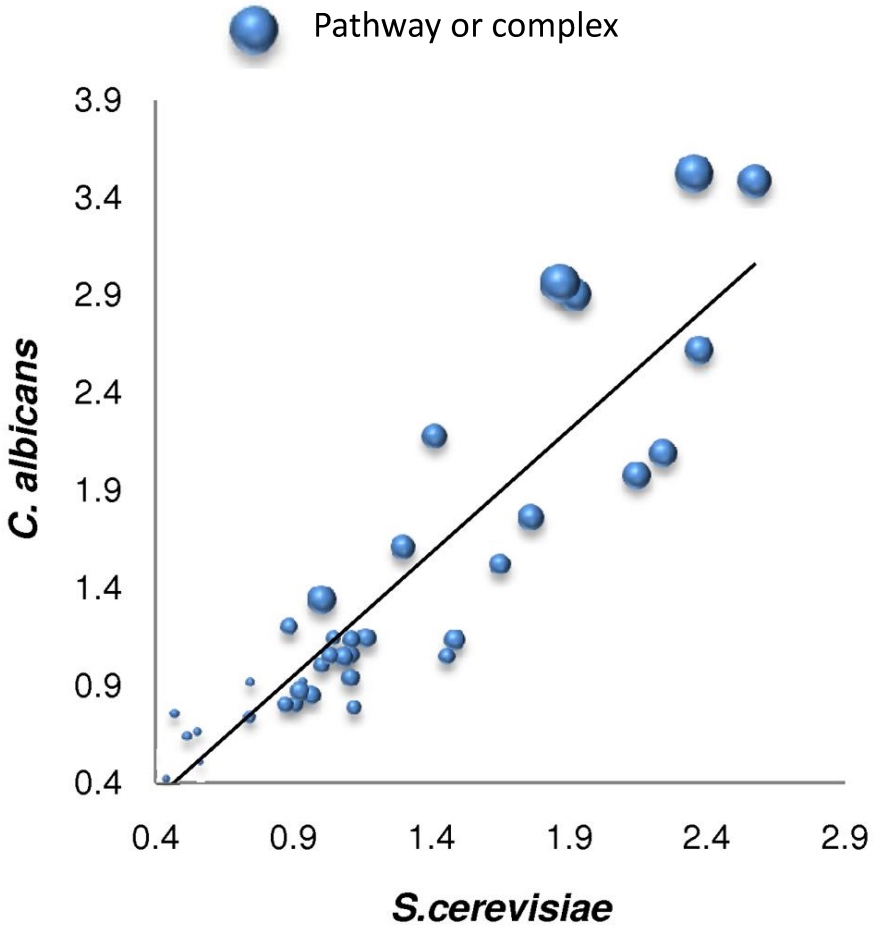
Phosphosites with known function are more likely conserved

S. cerevisiae phosphosites



It is possible that many sites have no biological function

Phosphorylation of functional ‘modules’ is well conserved



Average number of phosphosites per protein
(normalized for proteome coverage)

It is also possible that different sites have the same function in different species.

Beltrao et al. PLoS Bio 2009

Phosphosite conservation and protein abundance

- Some fraction of phosphosites are not functional
- Protein abundances span different orders of magnitude

Question: Are phosphosites more conserved in proteins of high or low abundance ?

Levy et al. Phil. Trans. R. Soc. B 2012

Evolution of kinase-substrate interactions

In order to study the evolution of kinase-substrate interactions. We need to predict what kinases are responsible for the phosphorylation.

How would you predict kinase-site interactions ?

Prediction of kinase-site interactions

In vitro

+

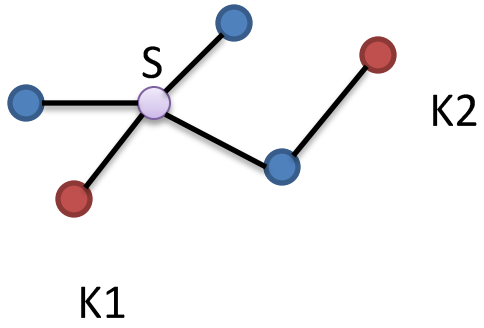
Co-expression
Co-localization
Scafolding
…

In vivo

Linding et al. Cell 2007

Prediction of kinase-site interactions

+

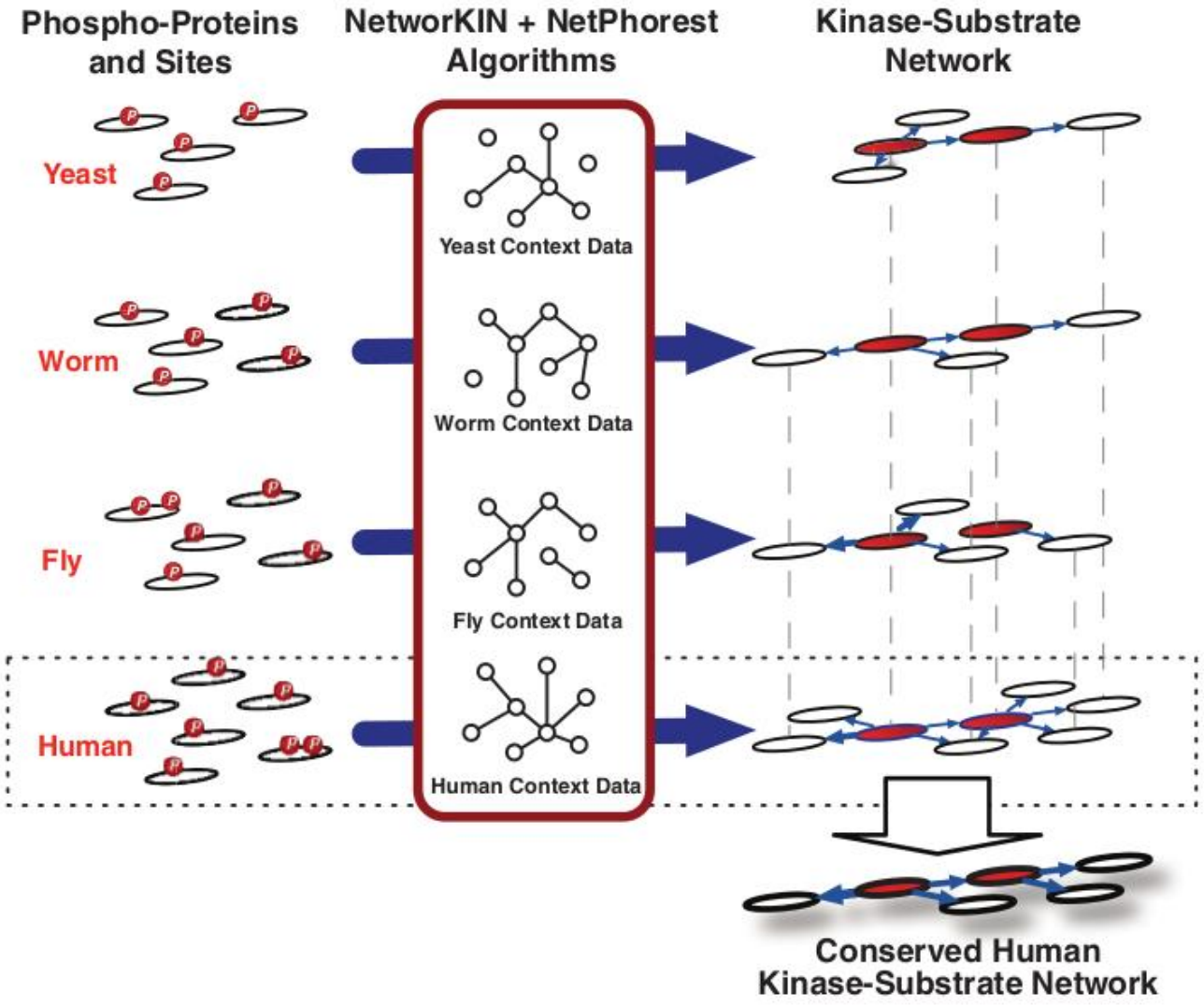


Models
from
Netphorest

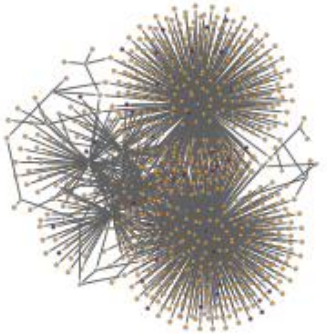
Functional
interactions
from
STRING

Linding et al. Cell 2007

Evolution of kinase-substrate interactions



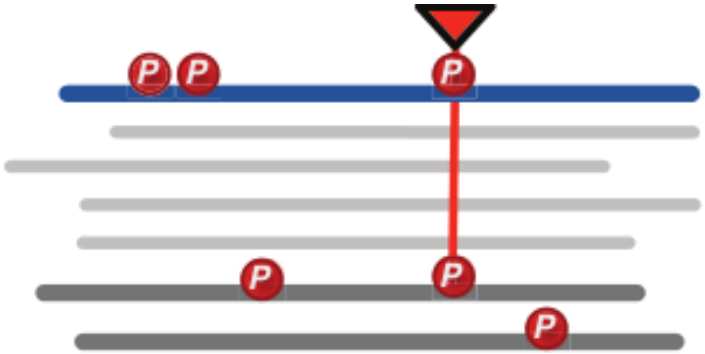
human kinase-substrate network
(113 kinases, 5515 substrates, 25,563 edges)



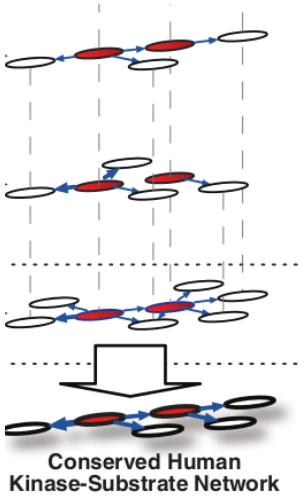
conserved ($k > 0$) kinase-substrate network
(27 kinases, 778 substrates, 1255 edges)

Tan et al. Science Signaling 2009

Evolution of kinase-substrate interactions



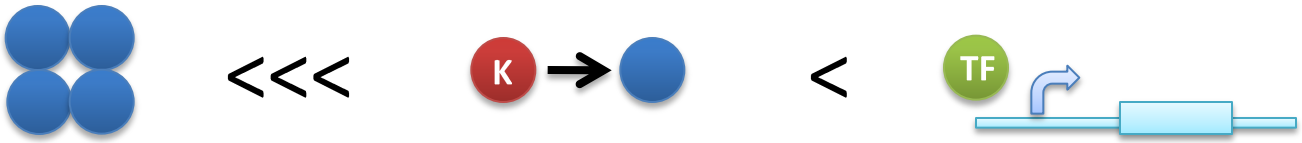
Vs.



- Only a small fraction of the phosphorylation sites are conserved in the same position of the alignment
- A significant fraction of the kinase-substrate interactions are conserved but with predicted changes in phosphosite position.
- The network of conserved kinase-substrate interactions is enriched in proteins that when mutated can result in disease.

Tan et al. Science Signaling 2009

Evolution of different types of physical interactions



<<<

<

Co-complex membership

Kinase-substrate interactions

TF-promoter interactions

The specificity of interaction determines the likelihood that a new interaction can be created by random variation. Lower specificity correlates with higher rate of divergence of interactions.

TF-gene Interactions diverge very quickly during evolution

	S. cerevisiae	*K. lactis*	*C. albicans*
	Fraction of genes mapped from species A bound in species B		
S. cerevisiae		42%	22%
K. lactis	16%		19%
C. albicans	7%	22%	

S. cerevisiae

K. lactis

C. albicans

ChIP-Chip of Mcm1 (TF) in 3 different species (100 to 400 My).

Very poor conservation (~20%) of interactions although binding site specificity appears conserved.

We don't know how much this change impacts on phenotypes.

Tuch et al. PLoS Biology 2008

Protein Complexes are largely conserved

OPEN ACCESS Freely available online

PLoS COMPUTATIONAL BIOLOGY

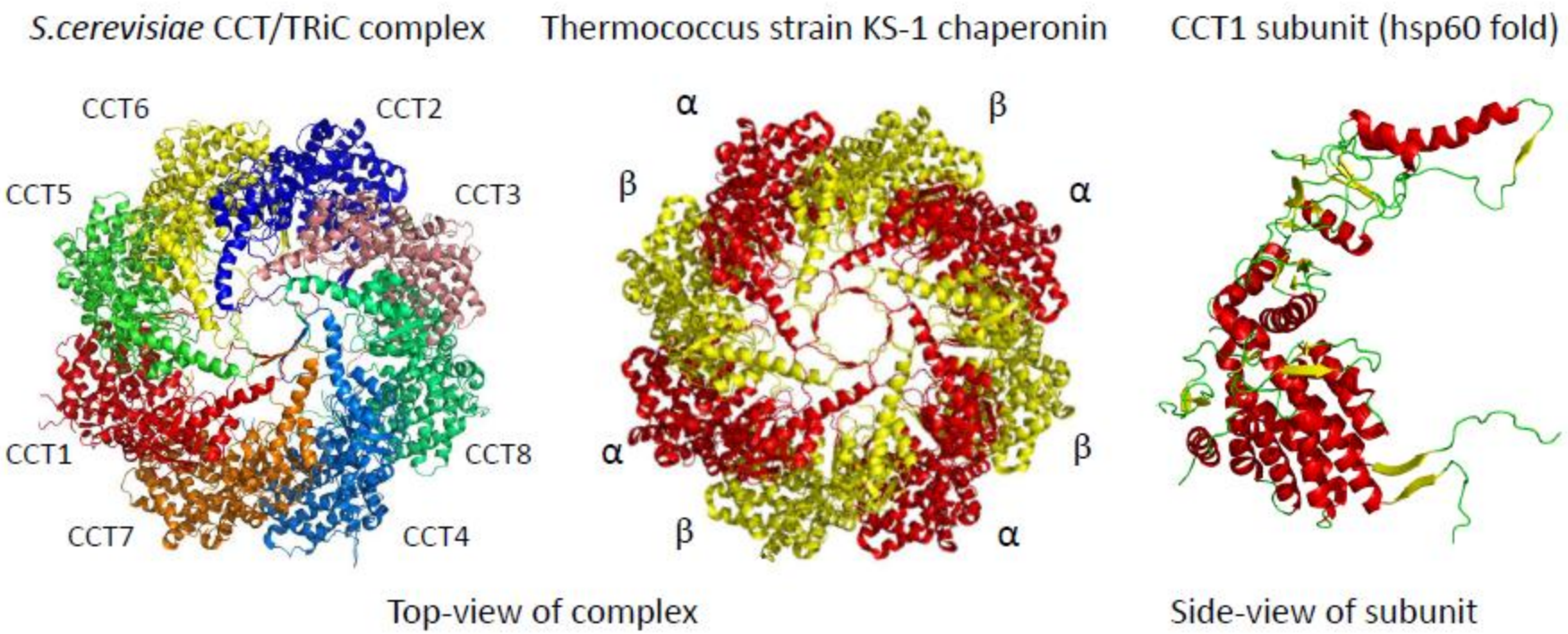
Protein Complex Evolution Does Not Involve Extensive Network Rewiring

2008

Teunis J. P. van Dam[1,2], Berend Snel[1,3]*

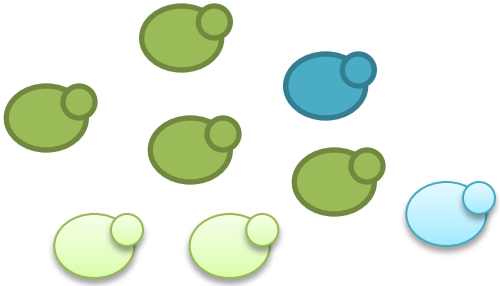
5,960 human co-complex pairs → 1,916 human co-complex pairs with orthologs in yeast → **90% of these are conserved**

Protein complex evolution by duplication-divergence

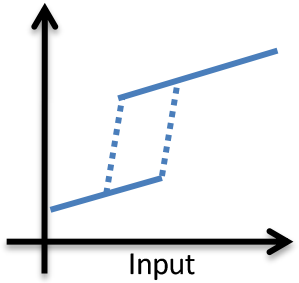


Pereira-Leal et al. Genome Biology 2007

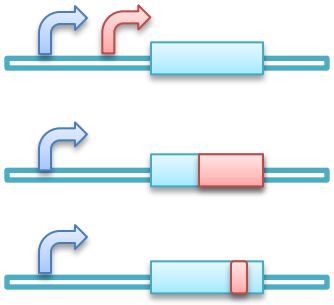
Fitness differences



Point mutations
Recombination
Duplications



Phenotypic changes



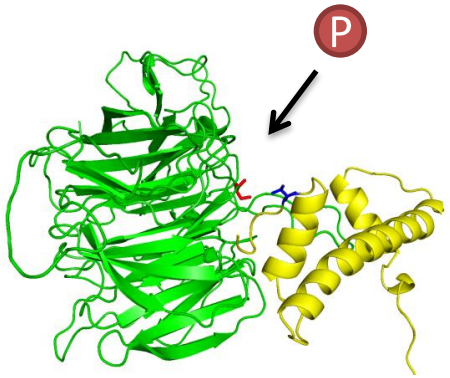
Changes in protein-protein,
protein-DNA interactions, etc

Systematic functional annotation of PTMs

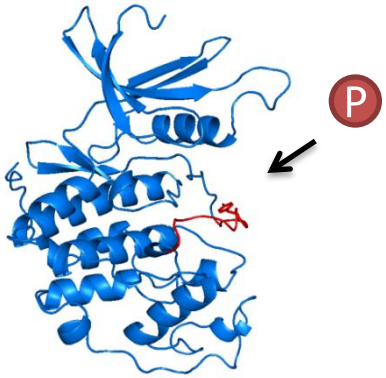
- PTMs diverge very quickly and it is likely that a significant number maybe have no function.
- To study the impact on fitness of these changes and for practical functional studies we need computational ways to predict the ones with function
  - Conservation of site and prediction of kinase (as described above)
  - **Assign a mechanist role for PTM**

Systematic functional annotation of PTMs

Globular regions

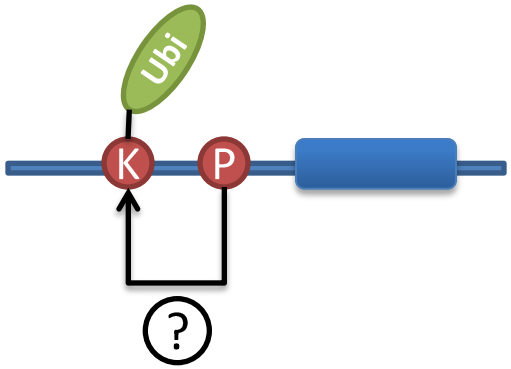


Regulating interfaces



Regulating domain activity (eg. allostery)

Unstructured regions



Phosphorylation switches

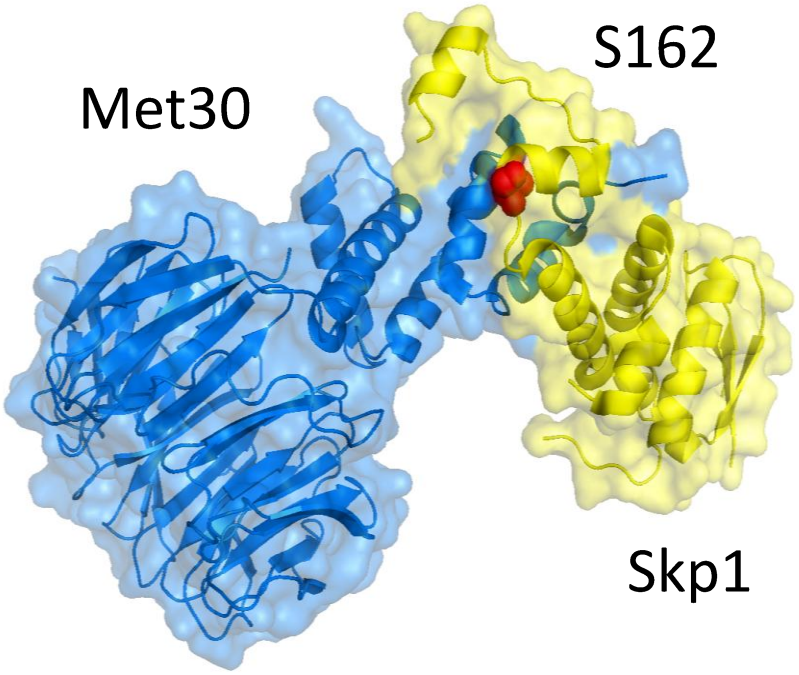
~25%	Phosphorylation	~75%
~45%	Acek/Ubi	~55%

Beltrao et al. Cell 2012

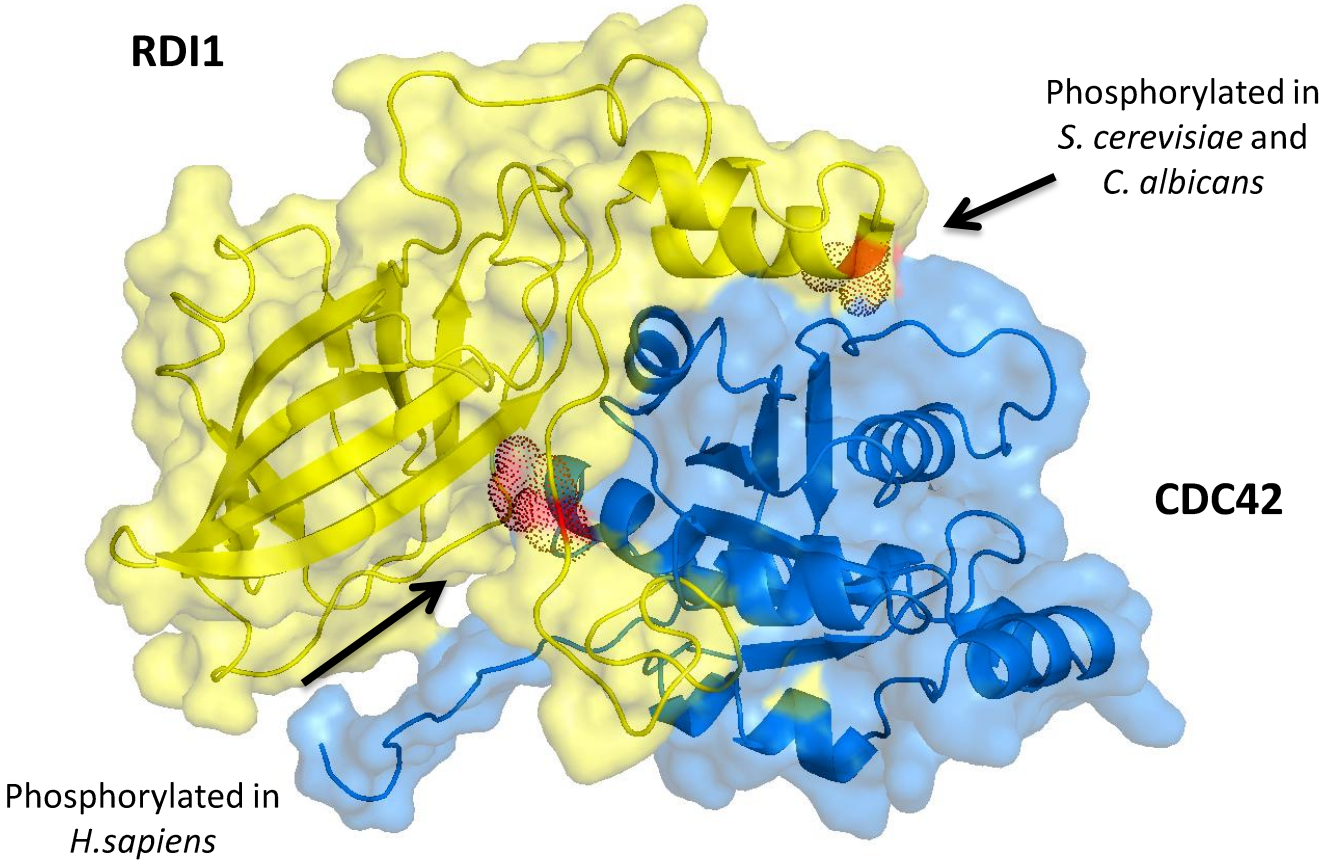
PTMs at interface residues

Interface residues were predicted based on xray, homology models, docking solutions and domain-domain contacts from the 3DID database

Example:
Skp1:Met30



Conservation of interface PTMs vs. conservation of function



Phosphorylated residues are 20 Amino-acids apart in sequence

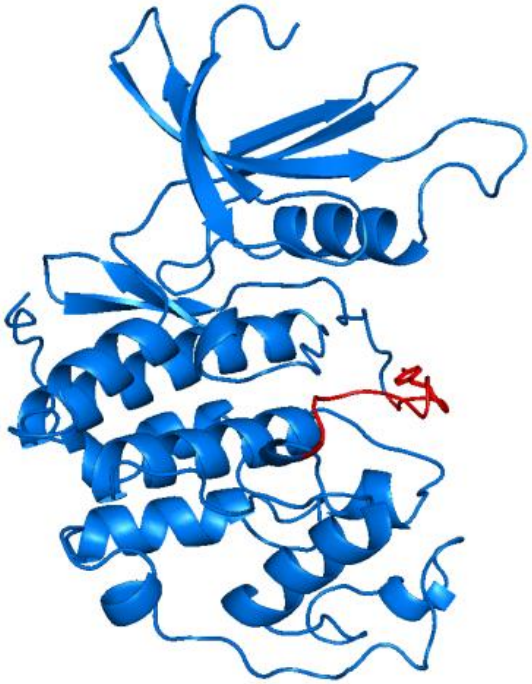
Conservation of function vs. conservation of phosphosites

Phosphosite	In *S. cerevisiae*	Conserved in another species by aln (+/- 5 pos)	Fraction conserved
All	22536	3585	**0.16**
In interface	401	142	**0.35**

	In *S. cerevisiae*	Conserved in another species	Fraction conserved
Interfaces with pSites (xray)	6	4	**0.66**
Interfaces with pSites (docking)	261	131	**0.50**
Interfaces with pSites (comparative modeling)	107	81	**0.76**

Positional conservation under-predicts conservation of function
It is possible to change a PTM site without changing its function

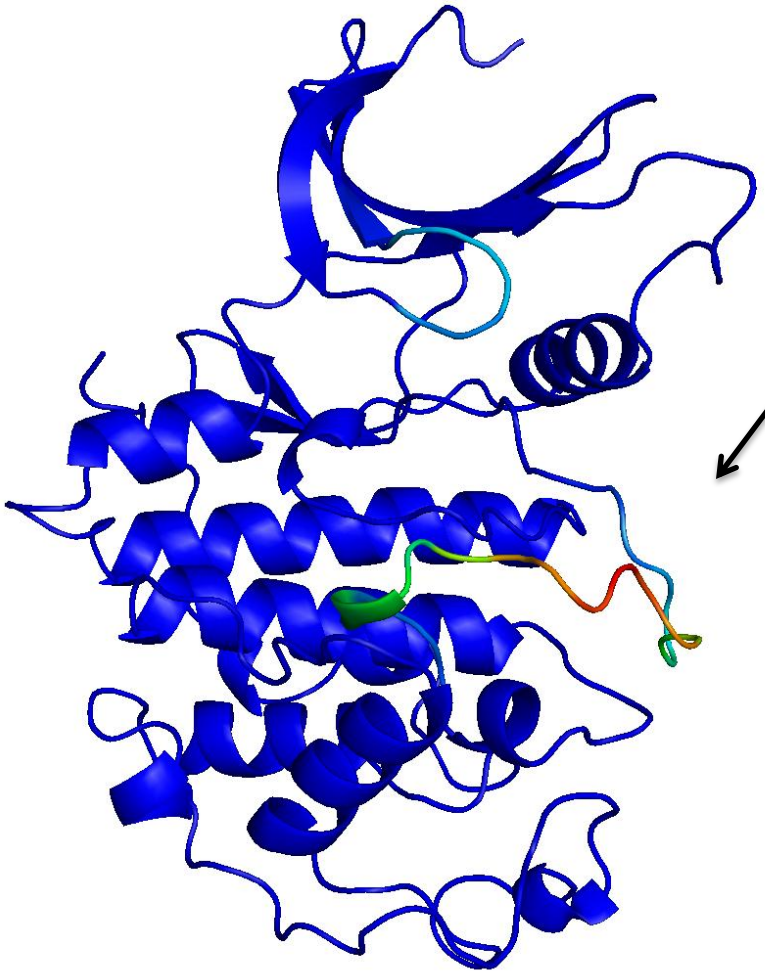
Regulation of domain activity



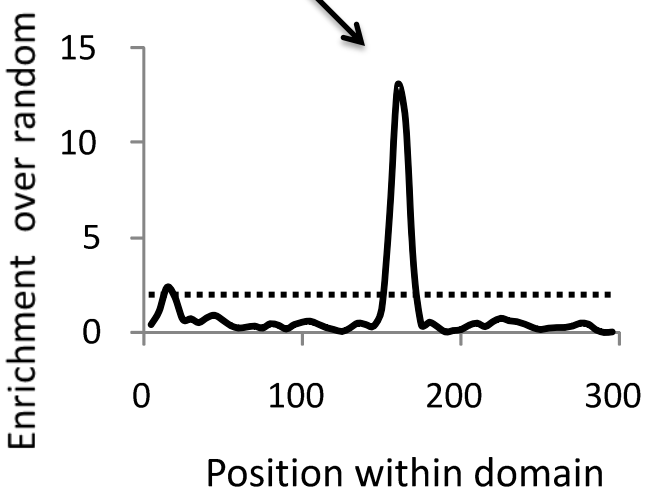
1. Pool all phosphosites (or other PTMs) and map them to representative structures of domain families

2. Look for sequence / structural enrichment (indicative of conservation)

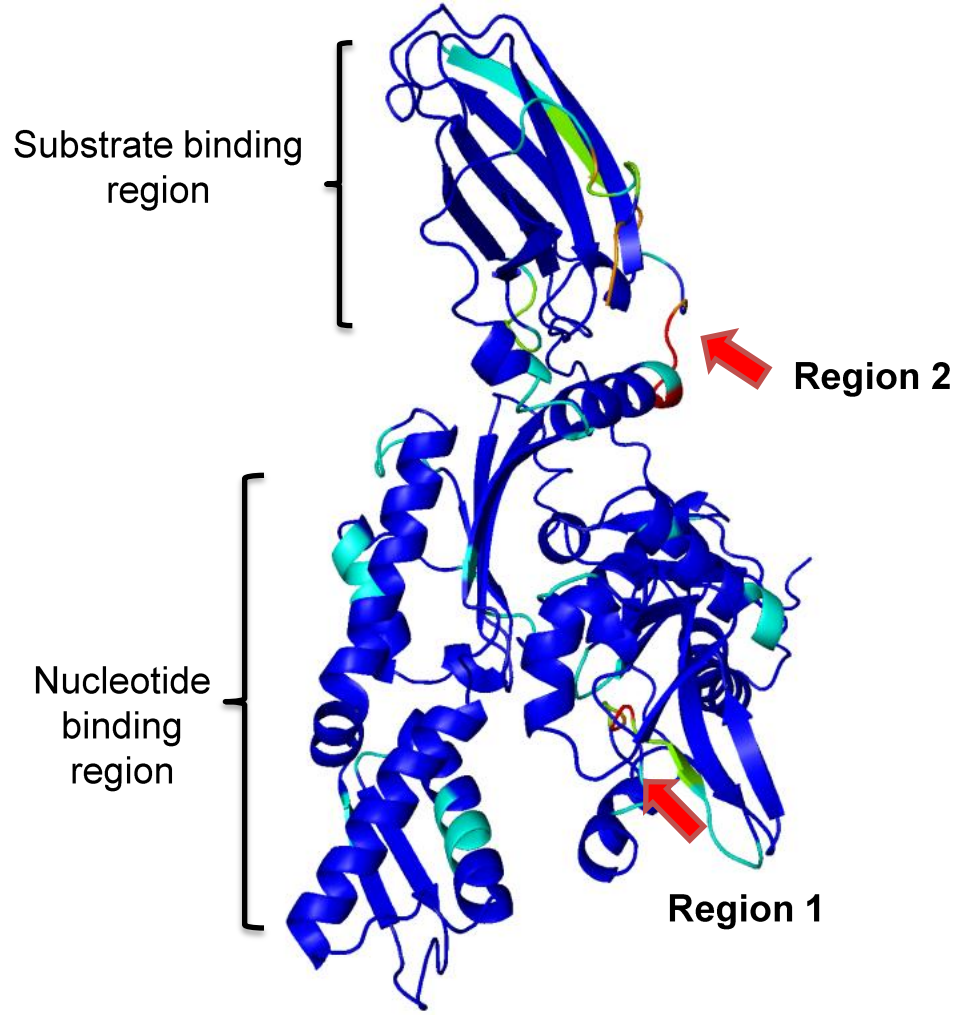
Protein kinase domain phosphorylation propensity

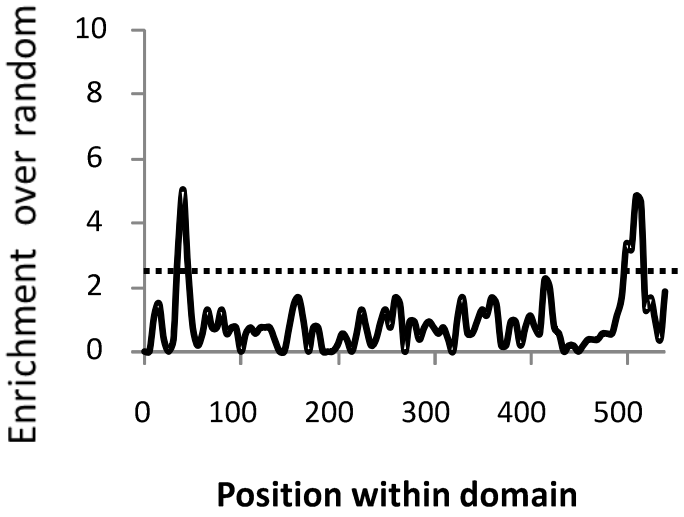


Activation loop (50% of psites)
Regulatory 'hot-spot'

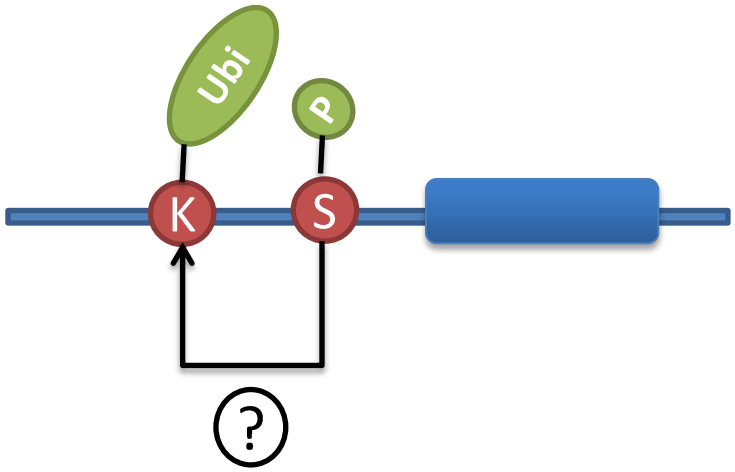


HSP70 domain Phosphorylation propensity





Co-regulation between different PTMs



We showed that:

- Different PTMś tend to cluster within proteins.
- Phosphosites near lysine modifications are more conserved than average phosphosites.

- The most promising but also the most challenging aspect of this analysis.
- New mass-spec and computational methods are needed (ex. dynamics after perturbation).

Recap - Evolution of targets

- PTM sites diverge very quickly at a rate that might be similar to the divergence of TF binding sites.
- Enzyme-protein interactions are more likely conserved than individual PTM positions
- PTMs with identifiable function shown an above average conservation
  - It is possible that a significant fraction have no function and it is possible that some sites change position without changing the function.
- Conservation of sites and interactions can be used to find important PTMs

Recap - Evolution of targets (Computational Methods)

- Standard comparative genomics has been used to study the evolution of PTM positions
- Machine learning methods have been used to predict kinase-protein interactions to study their evolution
- Structural bioinformatics has been useful to identify sites that are more likely functionally important.
- A lot of work still needs to be done in ranking sites/interactions according to functional importance as well as assigning the most likely mechanistic work.