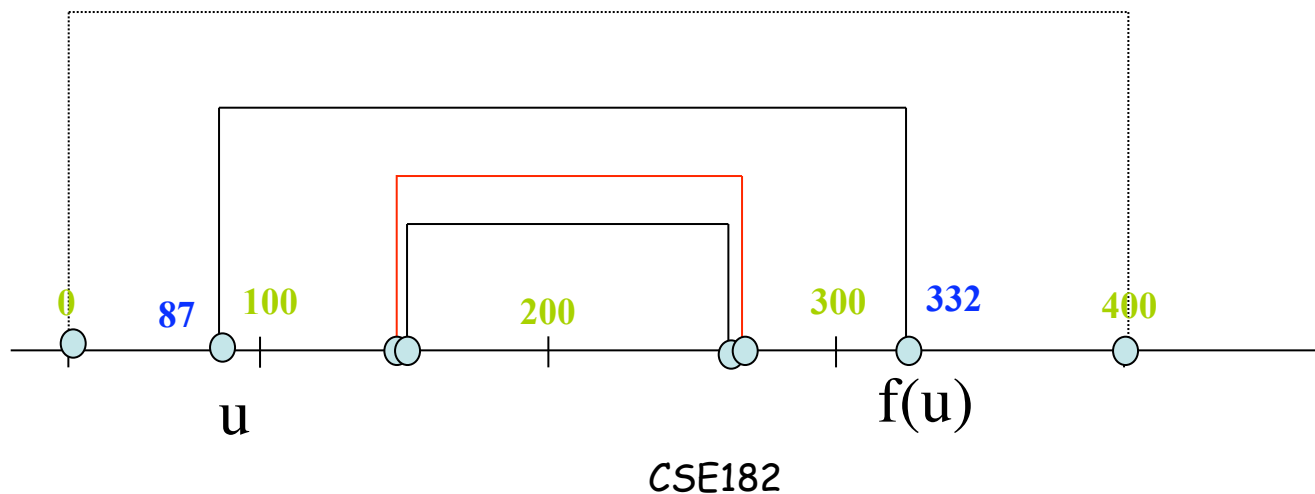


# CSE182-L13

Mass Spectrometry  
Quantitation and other applications

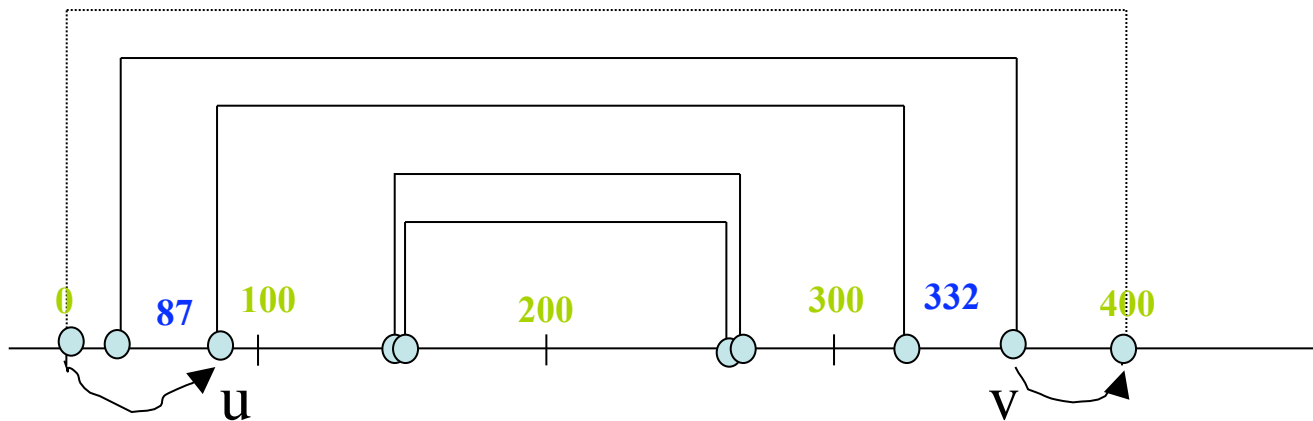
# The forbidden pairs method

- Sort the PRMs according to increasing mass values.
- For each node  $u$ ,  $f(u)$  represents the forbidden pair
- Let  $m(u)$  denote the mass value of the PRM.
- Let  $\delta(u)$  denote the score of  $u$
- Objective: Find a path of maximum score with no forbidden pairs.



# D.P. for forbidden pairs

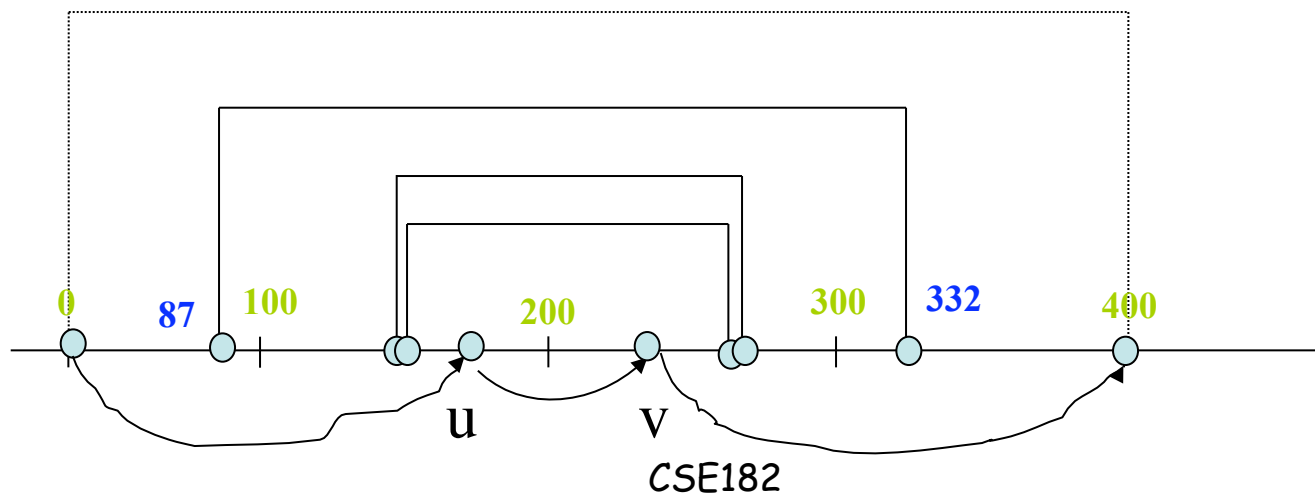
- Consider all pairs  $u, v$ 
  - $m[u] \leq M/2, m[v] > M/2$
- Define  $S(u, v)$  as the best score of a forbidden pair path from
  - $0 \rightarrow u$ , and  $v \rightarrow M$
- Is it sufficient to compute  $S(u, v)$  for all  $u, v$ ?



# D.P. for forbidden pairs

- Note that the best interpretation is given by

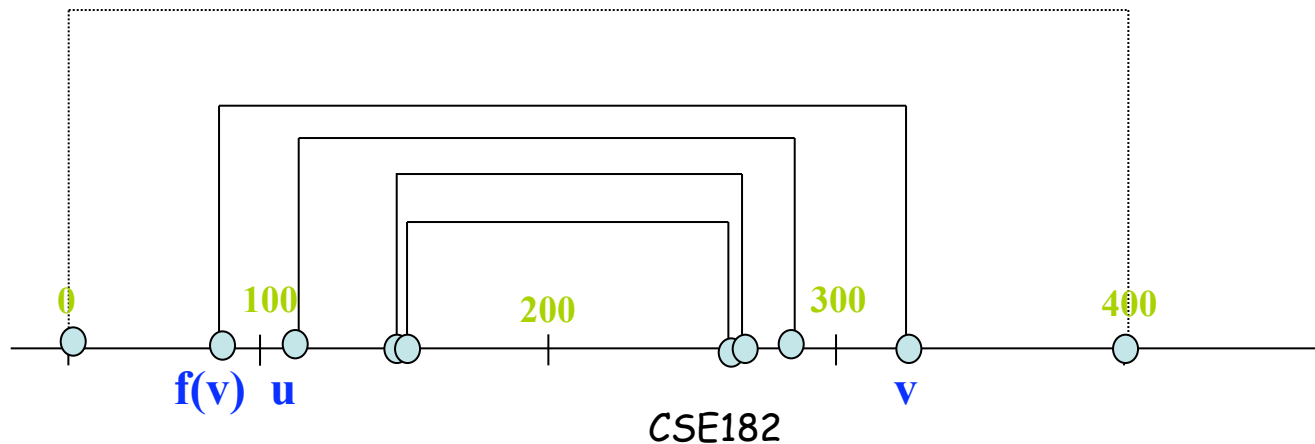
$$\max_{(u,v) \in E} S(u,v)$$



# D.P. for forbidden pairs

- Note that we have one of two cases.
  1. Either  $u > f(v)$  (and  $f(u) < v$ )
  2. Or,  $u < f(v)$  (and  $f(u) > v$ )
- Case 1.
  - Extend  $u$ , do not touch  $f(v)$

$$S(u, v) = \max_{\substack{u': (u', u) \in E \\ u' \neq f(v)}} S(u', v) + \delta(u)$$



# The complete algorithm

for all u /\*increasing mass values from 0 to M/2 \*/

for all v /\*decreasing mass values from M to M/2 \*/

if (u < f[v])

$$S[u,v] = \max_{\substack{(v,w) \in E \\ w \neq f(u)}} S[u,w] + \delta(v)$$

else if (u > f[v])

$$S[u,v] = \max_{\substack{(w,u) \in E \\ w \neq f(v)}} S[w,v] + \delta(u)$$

If (u,v) ∈ E

/\*maxI is the score of the best interpretation\*/

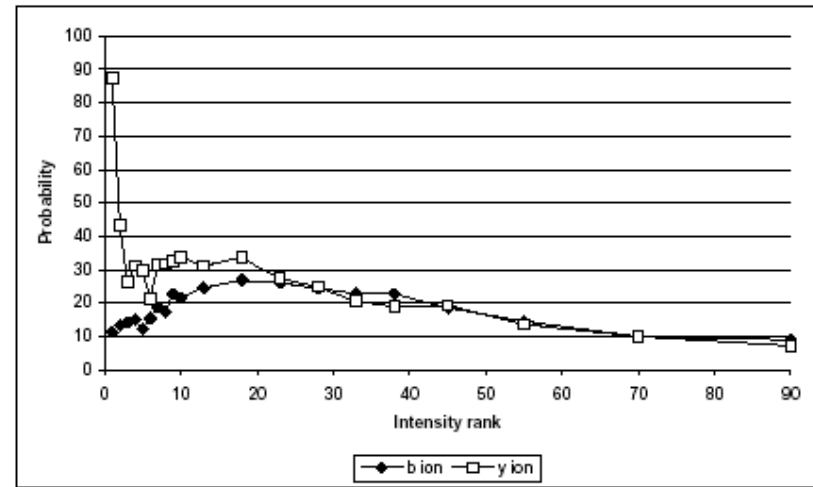
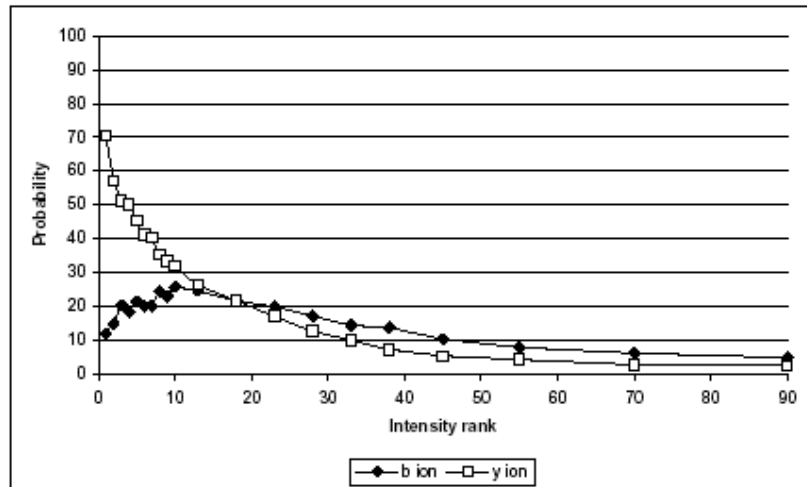
maxI = max {maxI, S[u,v]}

# De Novo: Second issue

- Given only b,y ions, a forbidden pairs path will solve the problem.
- However, recall that there are MANY other ion types.
  - Typical length of peptide: 15
  - Typical # peaks? 50-150?
  - #b/y ions?
  - Most ions are "Other"
    - a ions, neutral losses, isotopic peaks....

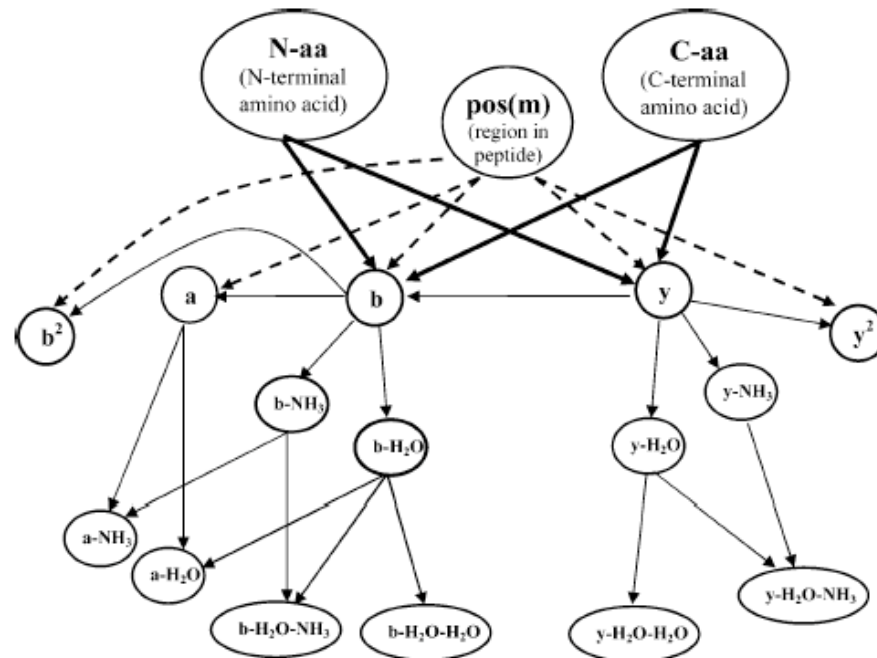
# De novo: Weighting nodes in Spectrum Graph

- Factors determining if the ion is b or y
  - Intensity (A large fraction of the most intense peaks are b or y)
  - Support ions
  - Isotopic peaks



# De novo: Weighting nodes

- A probabilistic network to model support ions (Peprnovo)



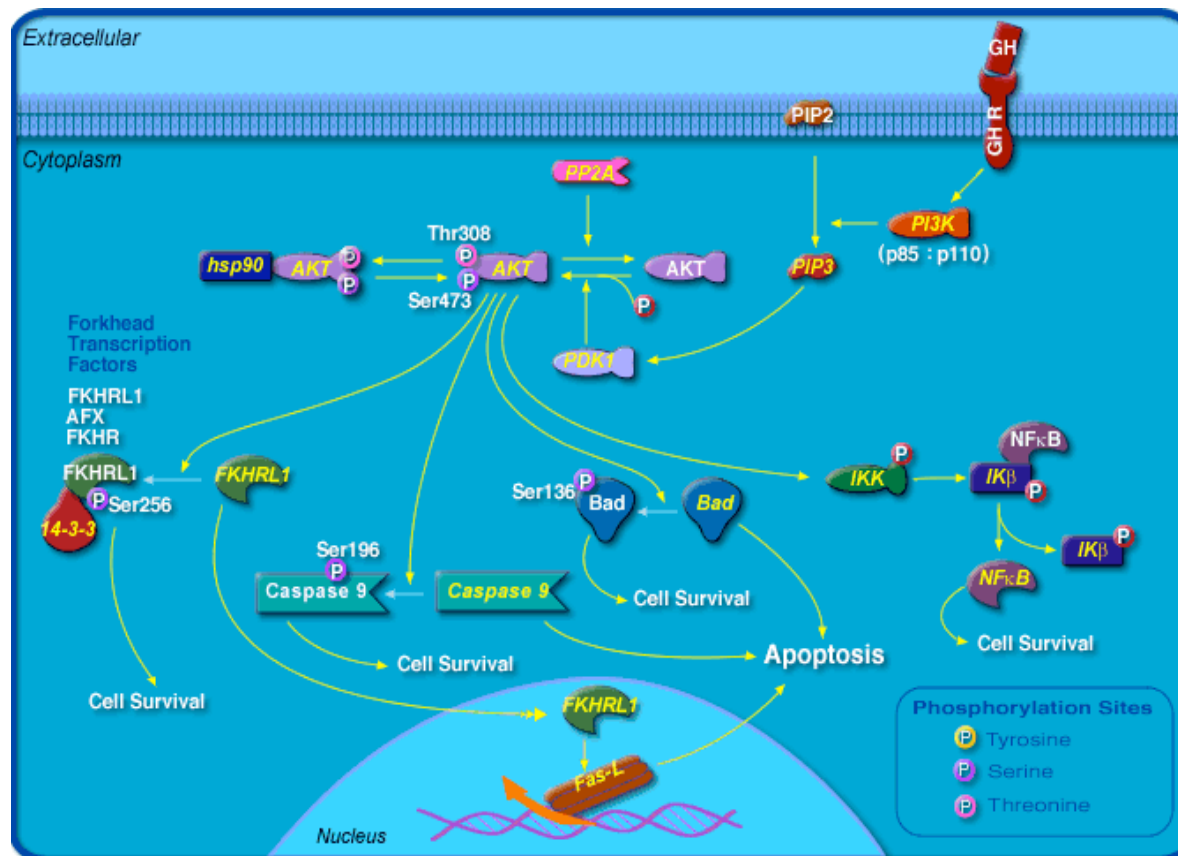
**Figure 1.** Probabilistic network for the CID fragmentation model of doubly charged tryptic peptides measured in an ion trap mass spectrometer. Three different types of relations are modeled in this network: (1) correlations between fragment ions (regular arrows); (2) dependencies due to the relative position of the cleavage site in the peptide (dashed arrows); (3) influence of flanking amino acids to the cleavage site (bold arrows).

$$\text{Score}(m, S) = \log \frac{P_{\text{CID}}(\bar{I}|m, S)}{P_{\text{RAND}}(\bar{I}|m, S)}$$

# De Novo Interpretation Summary

- The main challenge is to separate b/y ions from everything else (weighting nodes), and separating the prefix ions from the suffix ions (Forbidden Pairs).
- As always, the abstract idea must be supplemented with many details.
  - Noise peaks, incomplete fragmentation
  - In reality, a PRM is first scored on its likelihood of being correct, and the forbidden pair method is applied subsequently.
- In spite of these algorithms, de novo identification remains an error-prone process. When the peptide is in the database, db search is the method of choice.

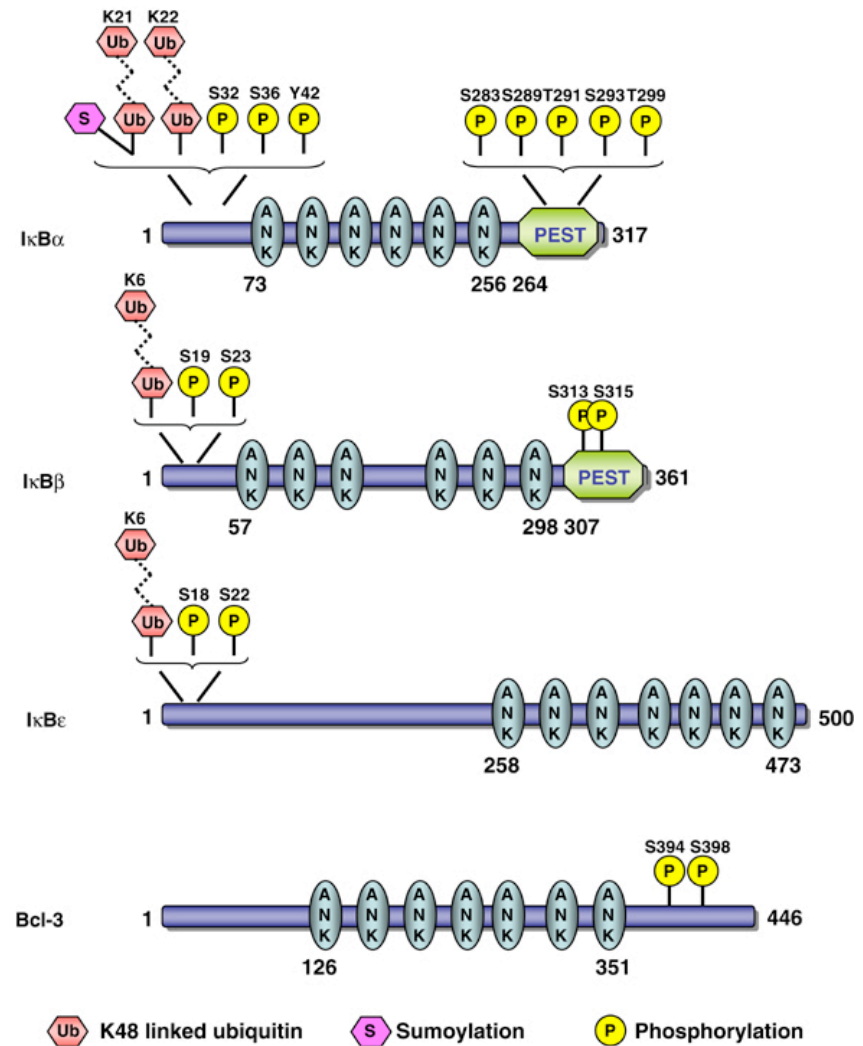
# The dynamic nature of the cell



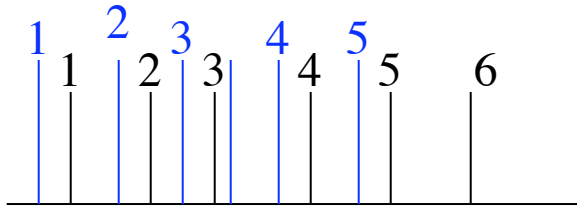
- The proteome of the cell is changing
- Various extra-cellular, and other signals activate pathways of proteins.
- A key mechanism of protein activation is PT modification
- These pathways may lead to other genes being switched on or off
- Mass Spectrometry is key to probing the proteome

# Post-translational modifications

- Post-translational modifications are key modulators of function.
- Usually, the PTM is created by attachment of a small chemical group



# What happens to the spectrum upon modification?



- Consider the peptide MSTYER.
- Either S, T, or Y (one or more) can be phosphorylated
- Upon phosphorylation, the b-, and y-ions shift in a characteristic fashion. Can you determine where the modification has occurred?

If T is phosphorylated,  $b_3$ ,  $b_4$ ,  $b_5$ ,  $b_6$ , and  $y_4$ ,  $y_5$ ,  $y_6$  will shift

# Effect of PT modifications on identification

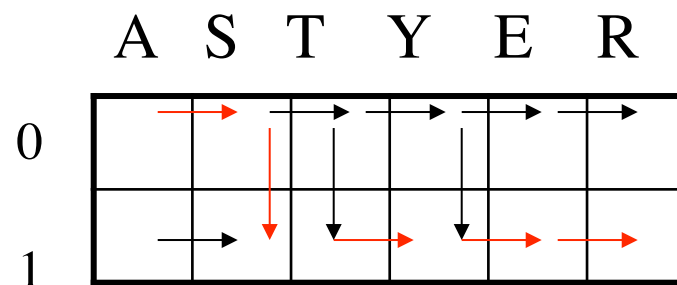
- The shifts do not affect de novo interpretation too much. Why?
- Database matching algorithms are affected, and must be changed.
- Given a candidate peptide, and a spectrum, can you identify the sites of modifications

# Db matching in the presence of modifications

- Consider MSTYER
- The number of modifications can be obtained by the difference in parent mass.
- With 1 phosphorylation event, we have 3 possibilities:
  - MS\*TYER
  - MST\*YER
  - MSTY\*ER
- Which of these is the best match to the spectrum?
- If 2 phosphorylations occurred, we would have 6 possibilities. Can you compute more efficiently?

# Scoring spectra in the presence of modification

- Can we predict the sites of the modification?
- A simple trick can let us predict the modification sites?
- Consider the peptide ASTYER. The peptide may have 0,1, or 2 phosphorylation events. The difference of the parent mass will give us the number of phosphorylation events. Assume it is 1.
- Create a table with the number of b,y ions matched at each breakage point assuming 0, or 1 modifications
- Arrows determine the possible paths. Note that there are only 2 downward arrows. The max scoring path determines the phosphorylated residue

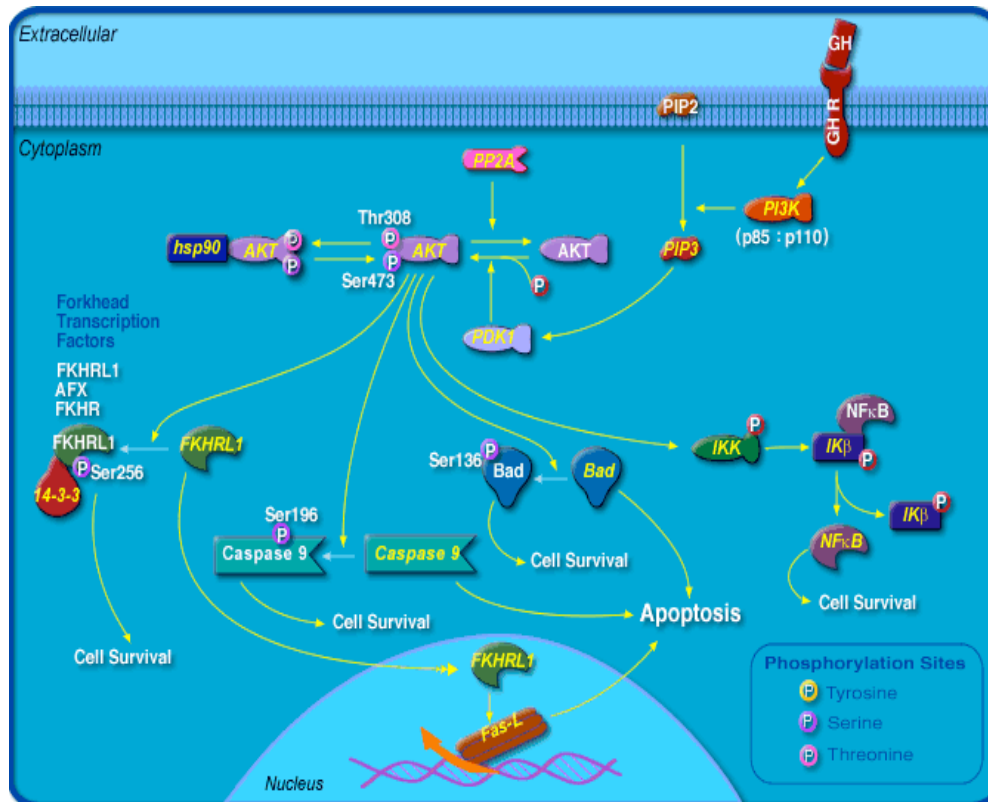


# Modifications Summary

- Modifications significantly increase the time of search.
- The algorithm speeds it up somewhat, but is still expensive

# MS based quantitation

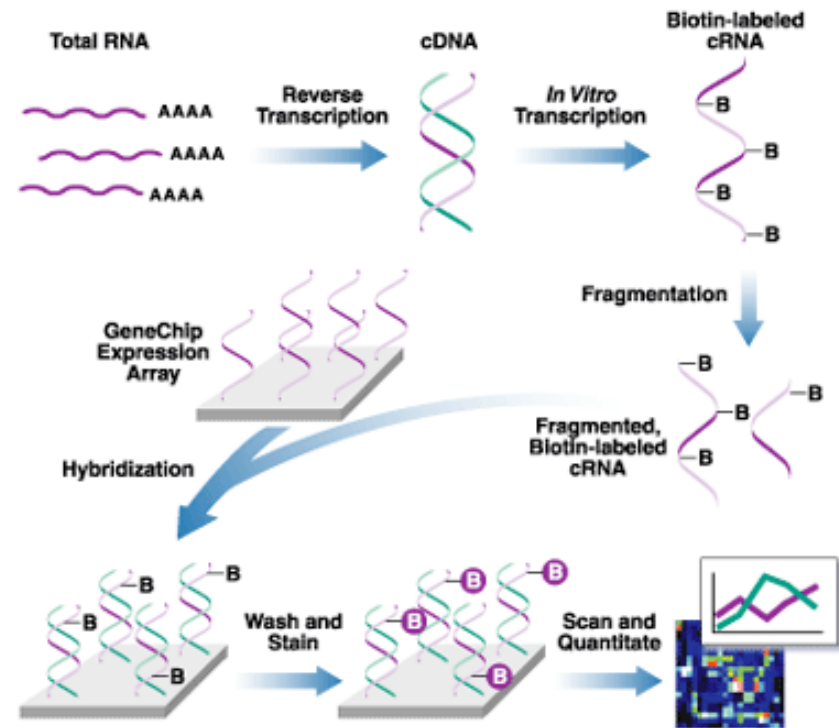
# The consequence of signal transduction



- The 'signal' from extra-cellular stimuli is transduced via phosphorylation.
- At some point, a 'transcription factor' might be activated.
- The TF goes into the nucleus and binds to DNA upstream of a gene.
- Subsequently, it 'switches' the downstream gene on or off

# Counting transcripts

- cDNA from the cell hybridizes to complementary DNA fixed on a 'chip'.
- The intensity of the signal is a 'count' of the number of copies of the transcript



# Quantitation: transcript versus Protein Expression

	Sample 1	Sample 2
mRNA1	100	20
mRNA1		
mRNA1		
mRNA1		
mRNA1		

	Sample 1	Sample2
Protein 1	35	4
Protein 2		
Protein 3		

Our Goal is to construct a matrix as shown for proteins, and RNA, and use it to identify differentially expressed transcripts/proteins

# Gene Expression

- Measuring expression at transcript level is done by micro-arrays and other tools
- Expression at the protein level is being done using mass spectrometry.
- Two problems arise:
  - Data: How to populate the matrices on the previous slide? ('easy' for mRNA, difficult for proteins)
  - Analysis: Is a change in expression significant? (Identical for both mRNA, and proteins).
- We will consider the data problem here. The analysis problem will be considered when we discuss micro-arrays.

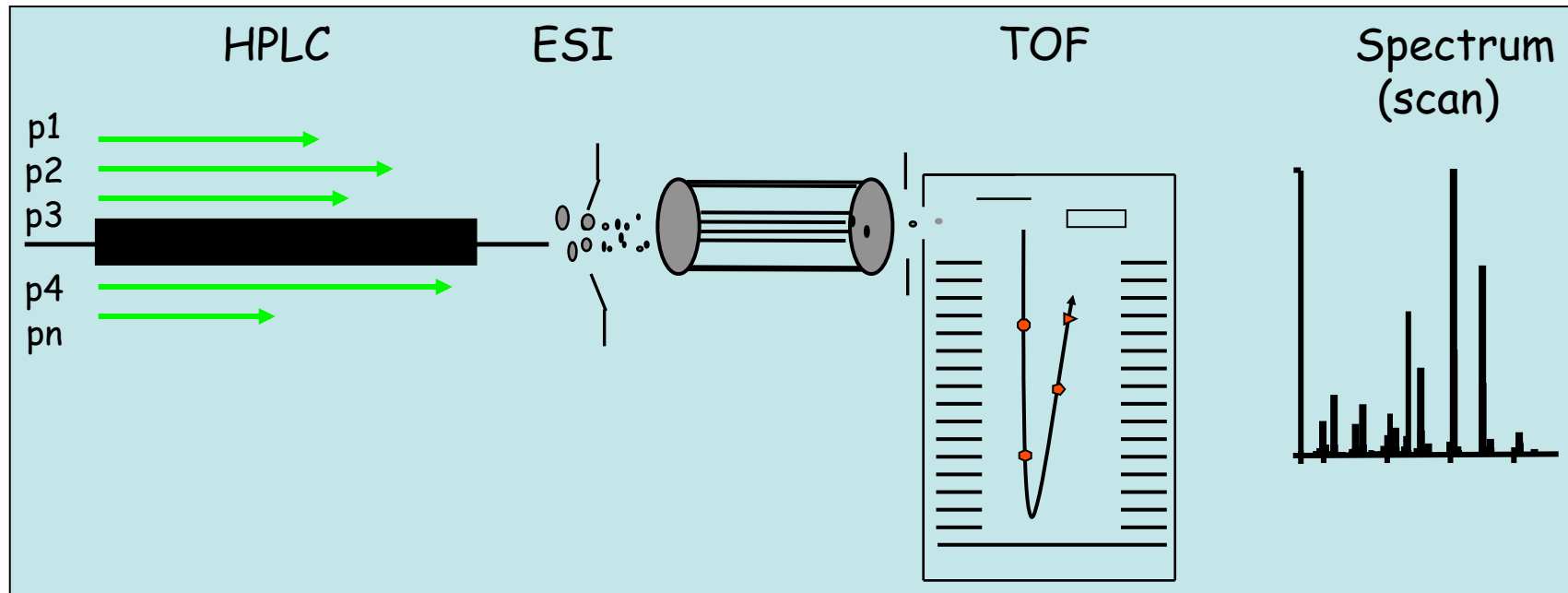
# MS based Quantitation

- The intensity of the peak depends upon
  - Abundance, ionization potential, substrate etc.
- We are interested in abundance.
- Two peptides with the same abundance can have very different intensities.
- **Assumption:** *relative* abundance can be measured by comparing the ratio of a peptide in 2 samples.

# Quantitation issues

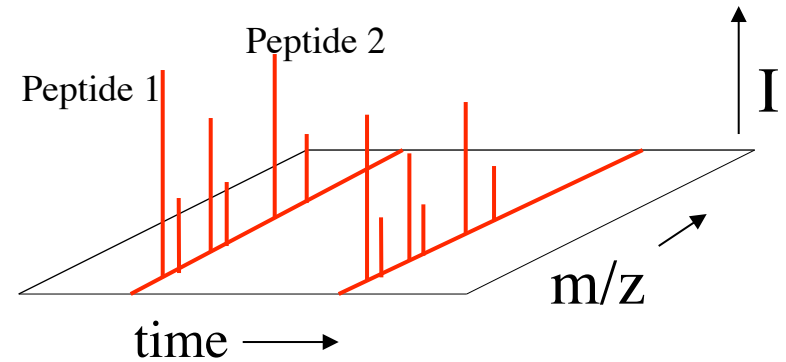
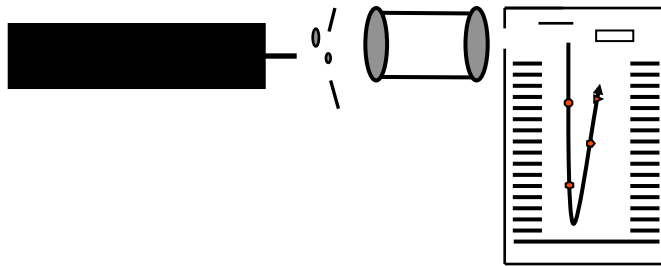
- The two samples might be from a complex mixture. How do we identify identical peptides in two samples?
- In micro-array this is possible because the cDNA is spotted in a precise location? Can we have a 'location' for proteins/peptides

# LC-MS based separation

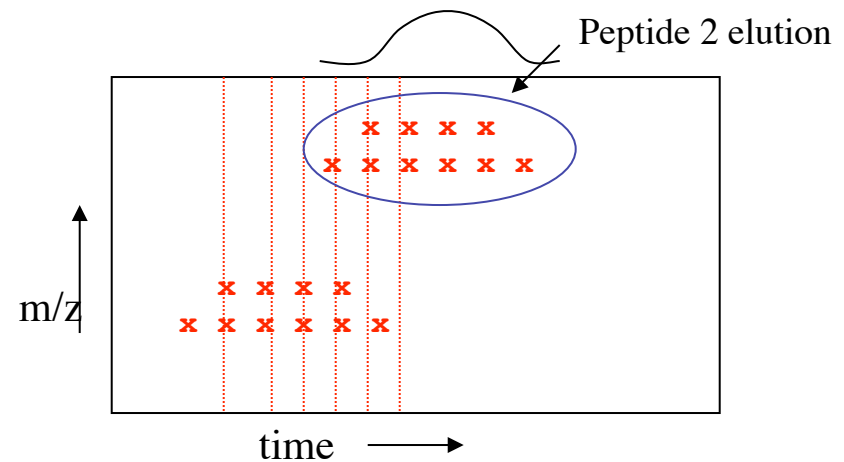


- As the peptides elute (separated by physiochemical properties), spectra is acquired.

# LC-MS Maps

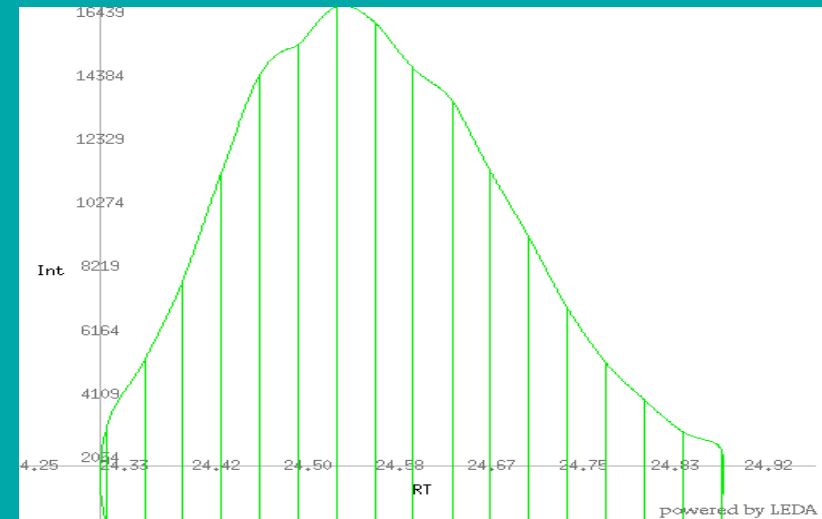
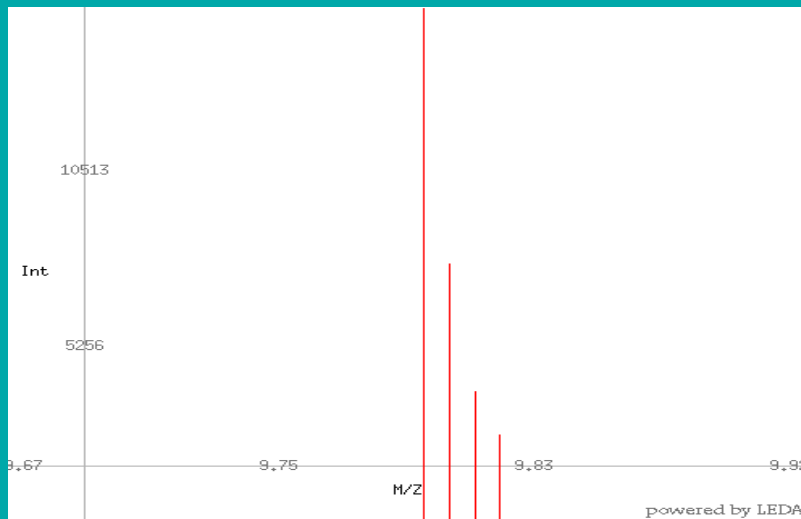
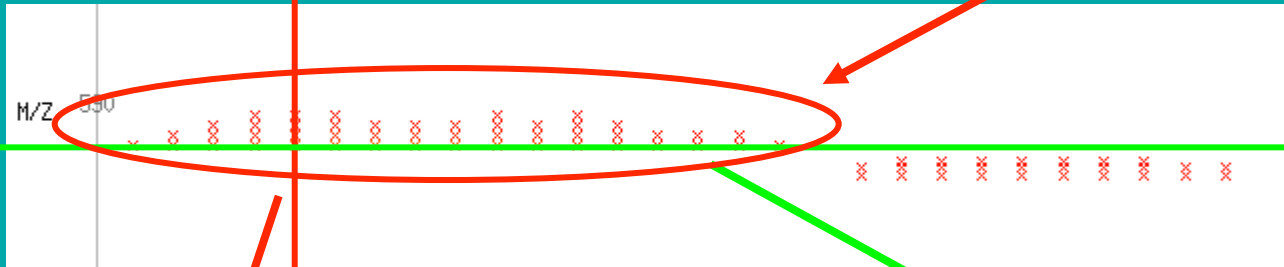


- A peptide/feature can be labeled with the triple (M,T,I):
  - monoisotopic M/Z, centroid retention time, and intensity
- An LC-MS map is a collection of features



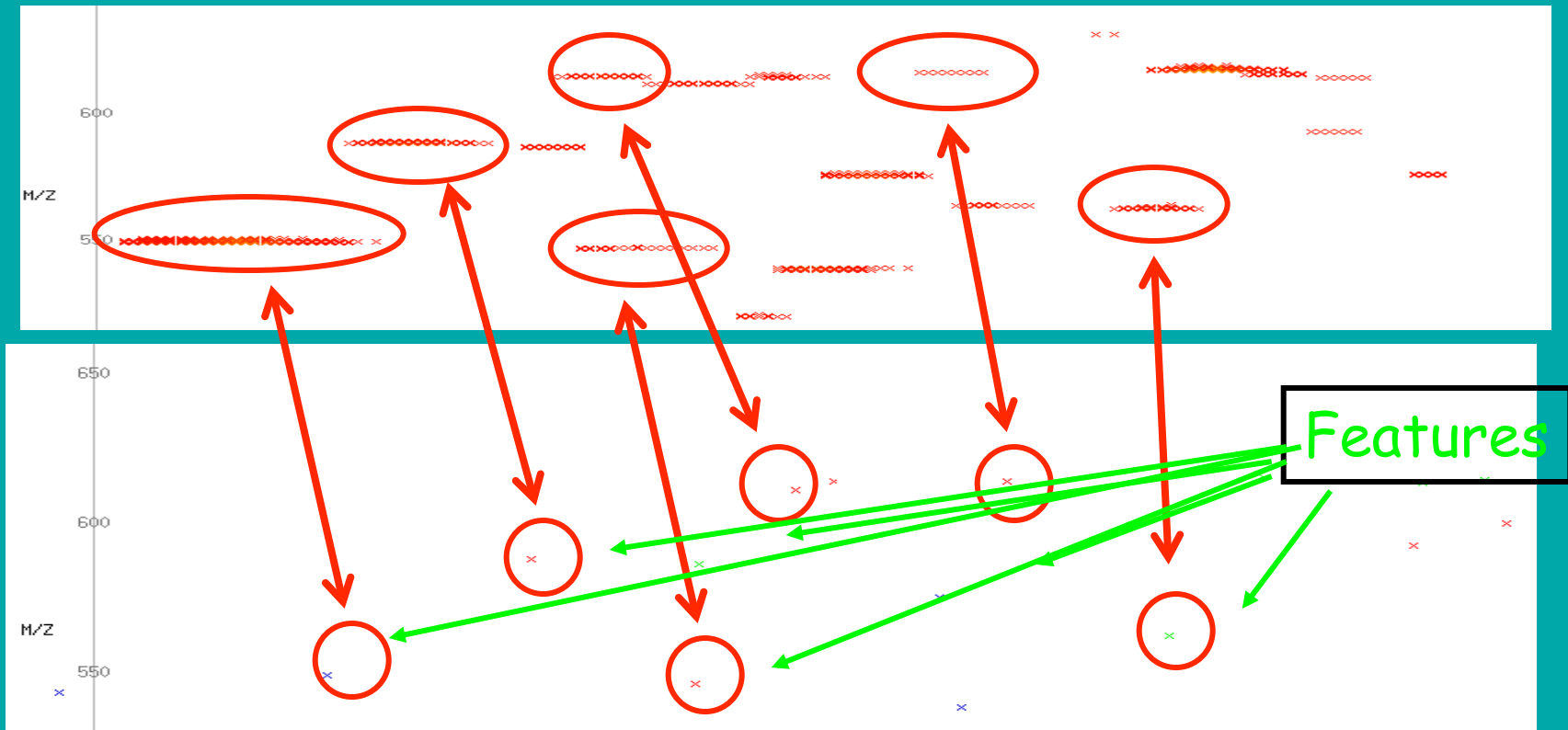
# Peptide Features

Peptide (feature)



Capture ALL peaks belonging to a peptide for quantification !

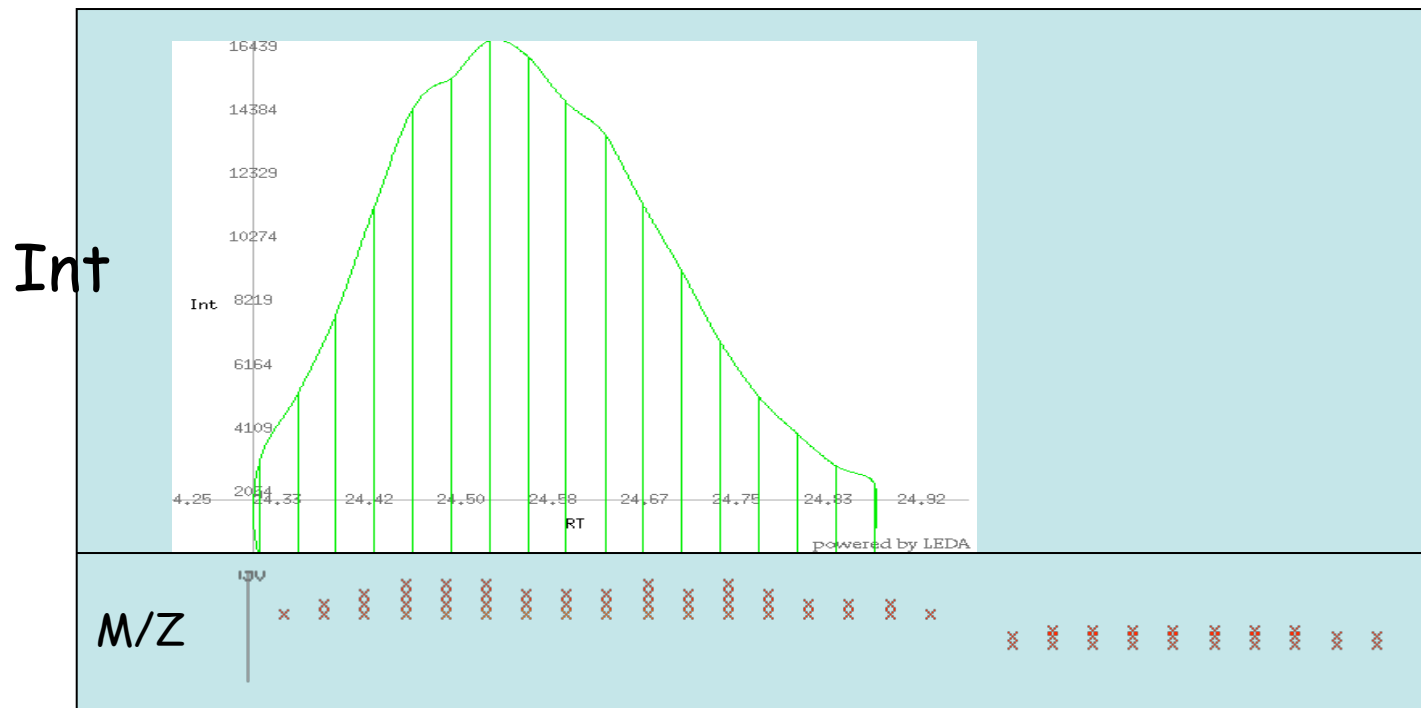
# Data reduction (feature detection)



- First step in LC-MS data analysis
- Identify 'Features': each feature is represented by
  - Monoisotopic  $M/Z$ , centroid retention time, aggregate intensity

# Feature Identification

- Input: given a collection of peaks (Time, M/Z, Intensity)
- Output: a collection of 'features'
  - Mono-isotopic m/z, mean time, Sum of intensities.
  - Time range  $[T_{beg}-T_{end}]$  for elution profile.
  - List of peaks in the feature.



# Feature Identification

- Approximate method:
- Select the dominant peak.
  - Collect all peaks in the same  $M/Z$  track
  - For each peak, collect isotopic peaks.
  - Note: the dominant peak is not necessarily the mono-isotopic one.

# Relative abundance using MS

- Recall that our goal is to construct an expression data-matrix with abundance values for each peptide in a sample. How do we identify that it is the same peptide in the two samples?
- Direct Map comparison
- Differential Isotope labeling (ICAT/SILAC)
- External standards (AQUA)

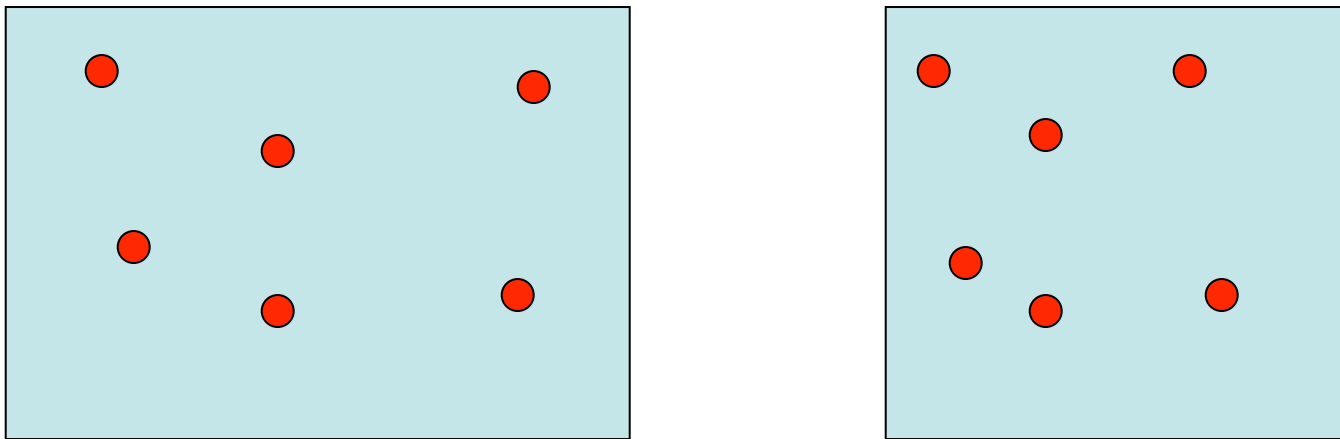
# Map Comparison for Quantification

Map 1 (normal)

Map 2 (diseased)



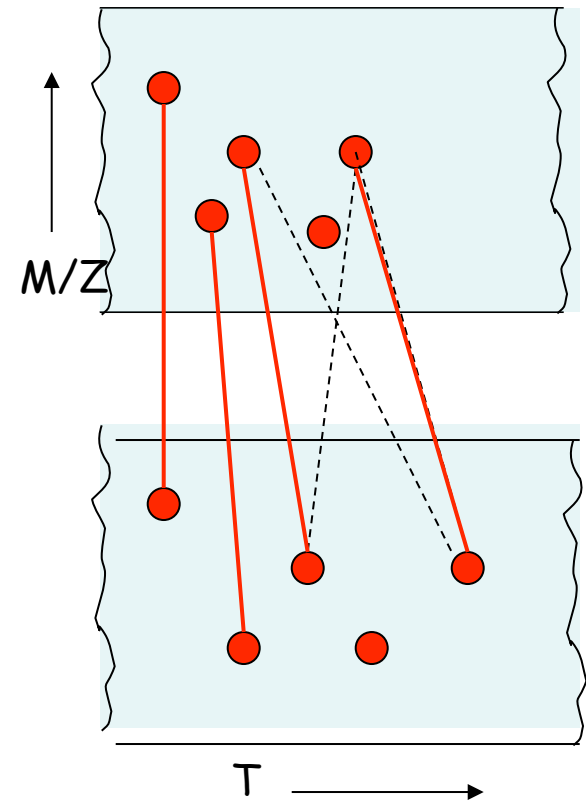
# Time scaling: Approach 1 (geometric matching)



- Match features based on  $M/Z$ , and (loose) time matching. Objective  $\sum_f (t_1 - t_2)^2$
- Let  $t_2' = a t_2 + b$ . Select  $a, b$  so as to minimize  $\sum_f (t_1 - t_2')^2$

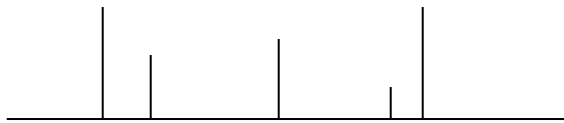
# Geometric matching

- Make a graph. Peptide  $a$  in LCMS1 is linked to all peptides with identical  $m/z$ .
- Each edge has score proportional to  $t_1/t_2$
- Compute a maximum weight matching.
- The ratio of times of the matched pairs gives  $a$ .
- Rescale and compute the scaling factor



# Approach 2: Scan alignment

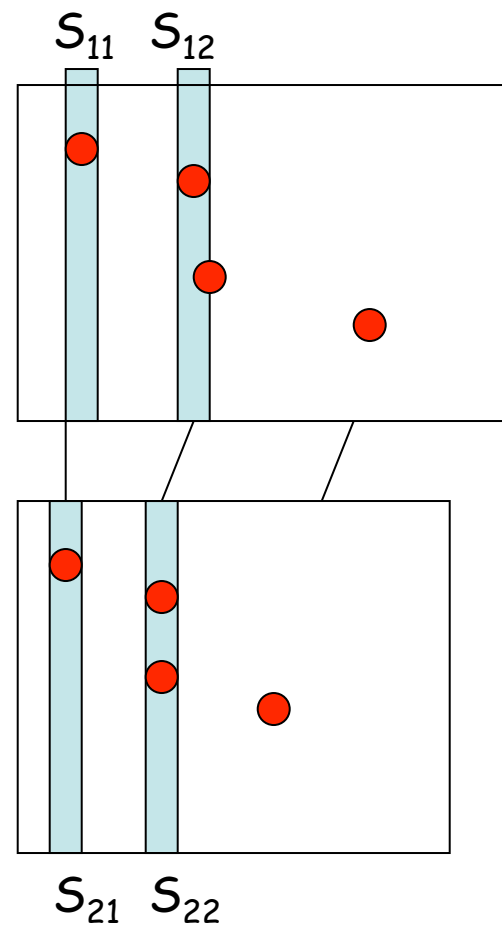
- Each time scan is a vector of intensities.
- Two scans in different runs can be scored for similarity (using a dot product)



$$S_{1i} = 10 \ 5 \ 0 \ 0 \ 7 \ 0 \ 0 \ 2 \ 9$$

$$S_{2j} = 9 \ 4 \ 2 \ 3 \ 7 \ 0 \ 6 \ 8 \ 3$$

$$M(S_{1i}, S_{2j}) = \sum_k S_{1i}(k) S_{2j}(k)$$

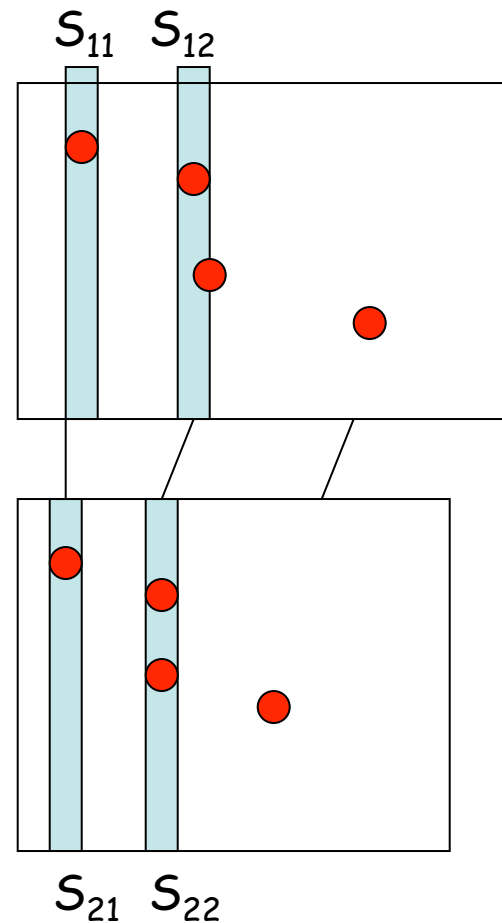


# Scan Alignment

- Compute an alignment of the two runs
- Let  $W(i,j)$  be the best scoring alignment of the first  $i$  scans in run 1, and first  $j$  scans in run 2

$$W(i,j) = \max \begin{cases} W(i-1,j-1) + M[S_{1i}, S_{2j}] \\ W(i-1,j) + \dots \\ W(i,j-1) + \dots \end{cases}$$

- Advantage: does not rely on feature detection.
- Disadvantage: Might not handle affine shifts in time scaling, but is better for local shifts



# Chemistry based methods for comparing peptides

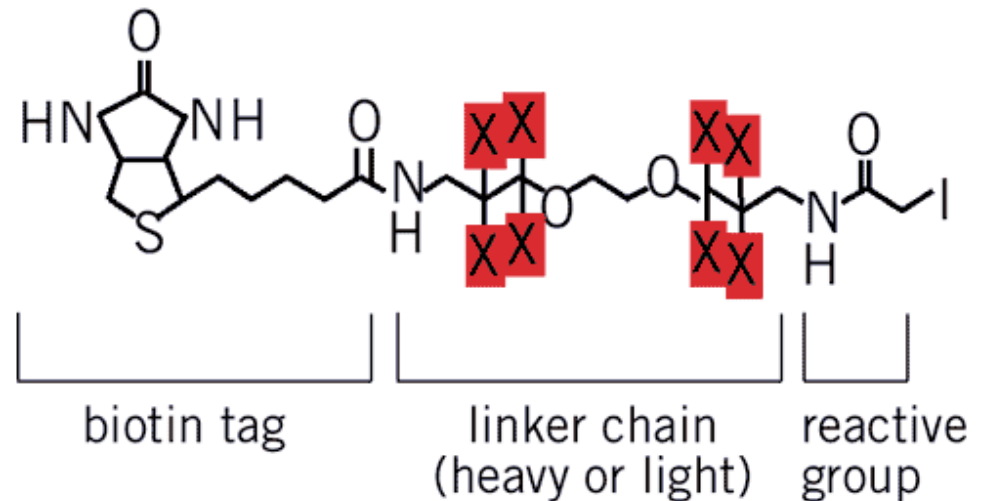
# ICAT

- The reactive group attaches to Cysteine
- Only Cys-peptides will get tagged
- The biotin at the other end is used to pull down peptides that contain this tag.
- The X is either Hydrogen, or Deuterium (Heavy)
  - Difference = 8Da

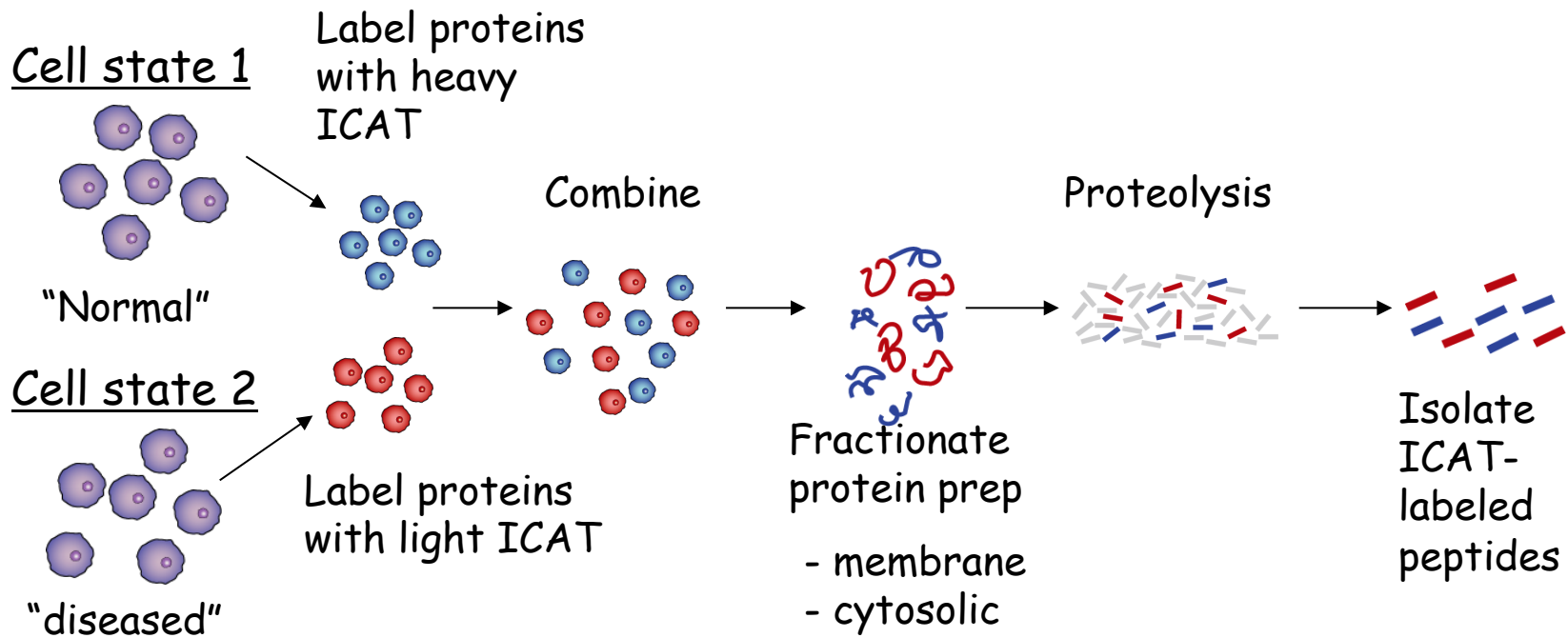
## Isotope-Coded Affinity Tags

heavy reagent: D8-ICAT Reagent (X=deuterium)

light reagent: D0-ICAT Reagent (X=hydrogen)



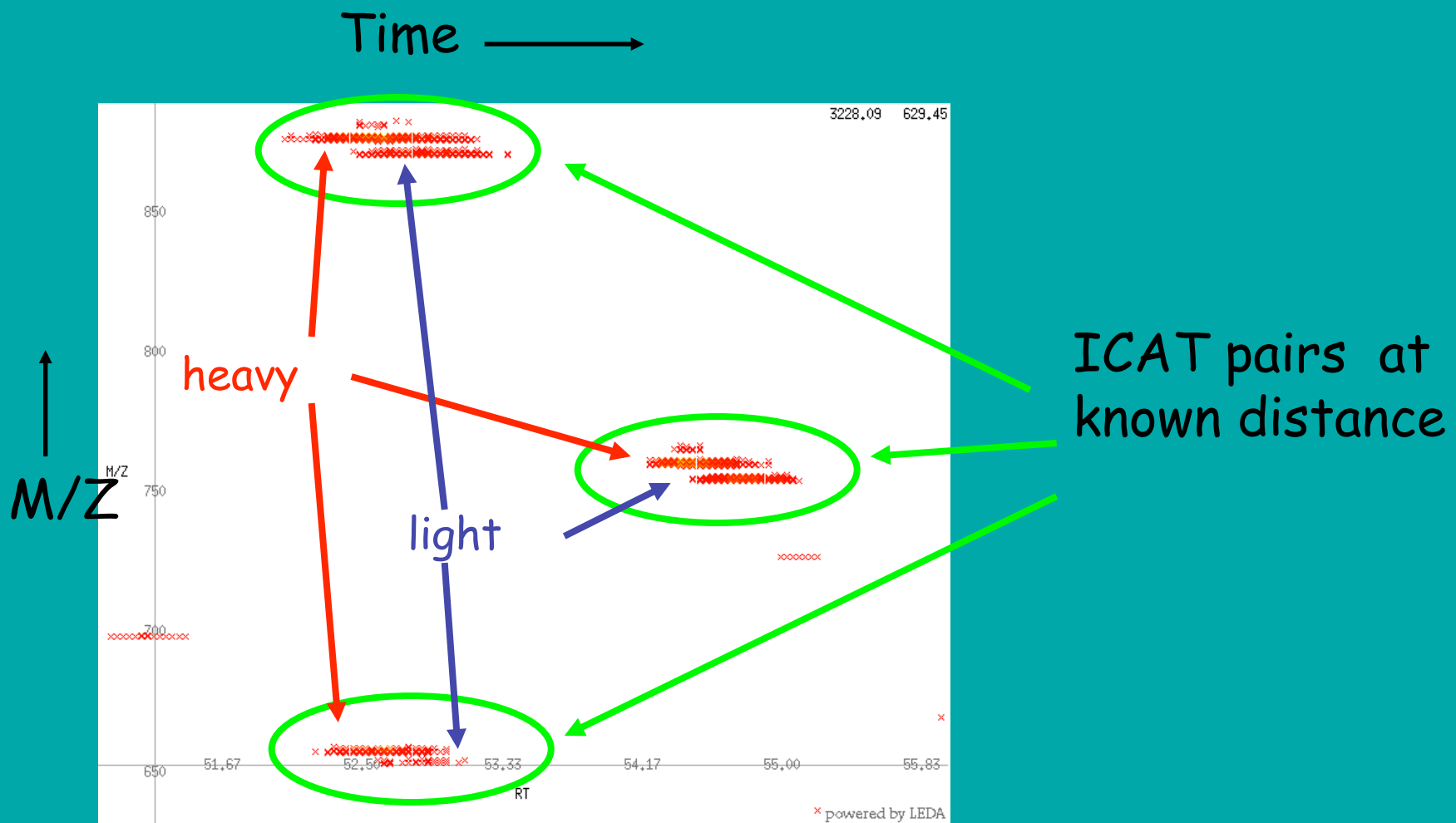
# ICAT



*Nat. Biotechnol.* 17: 994-999,1999

- ICAT reagent is attached to particular amino-acids (Cys)
- Affinity purification leads to simplification of complex mixture

# Differential analysis using ICAT

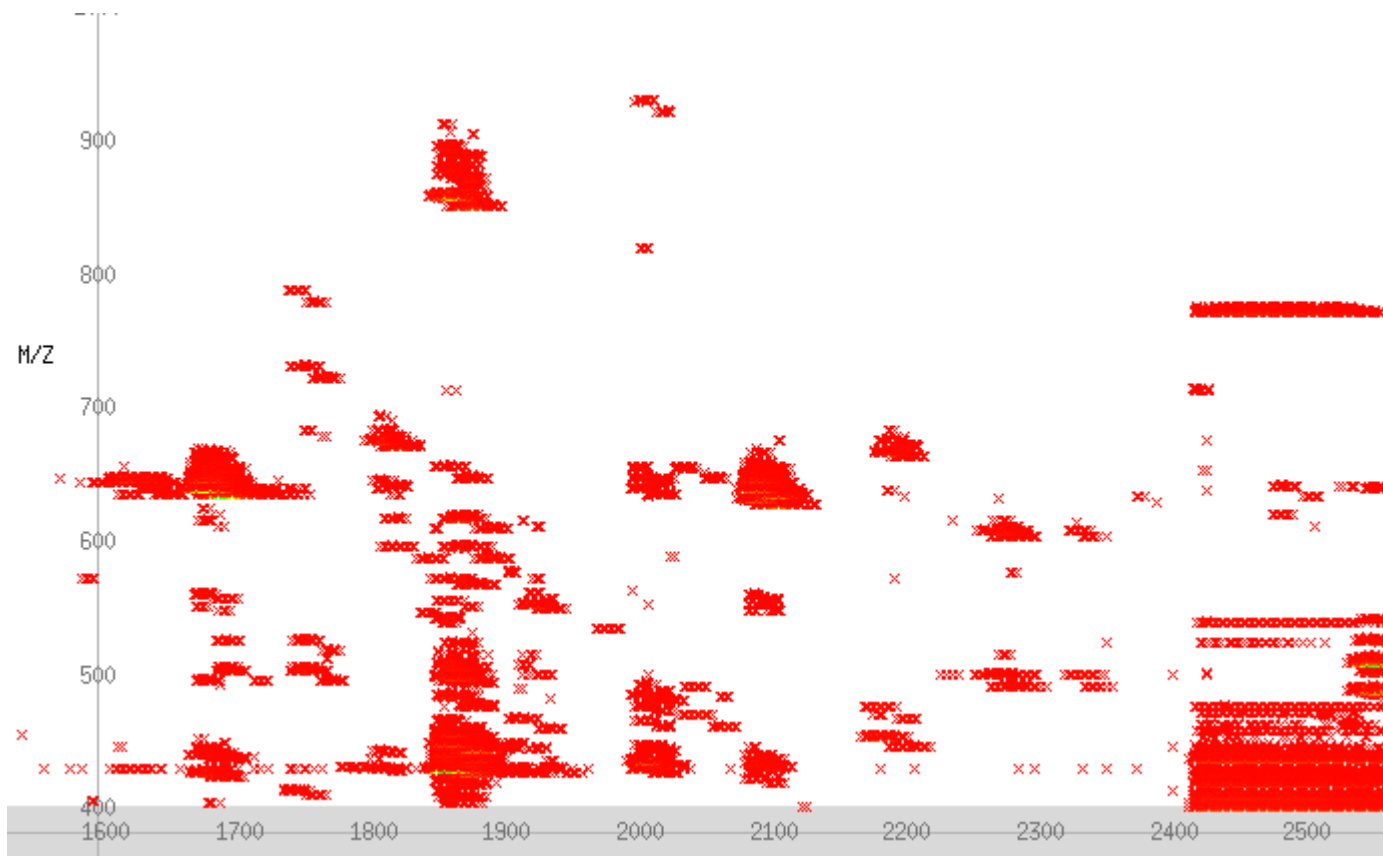


# ICAT issues

- The tag is heavy, and decreases the dynamic range of the measurements.
- The tag might break off
- Only Cysteine containing peptides are retrieved  
Non-specific binding to strepdavidin

# Serum ICAT data

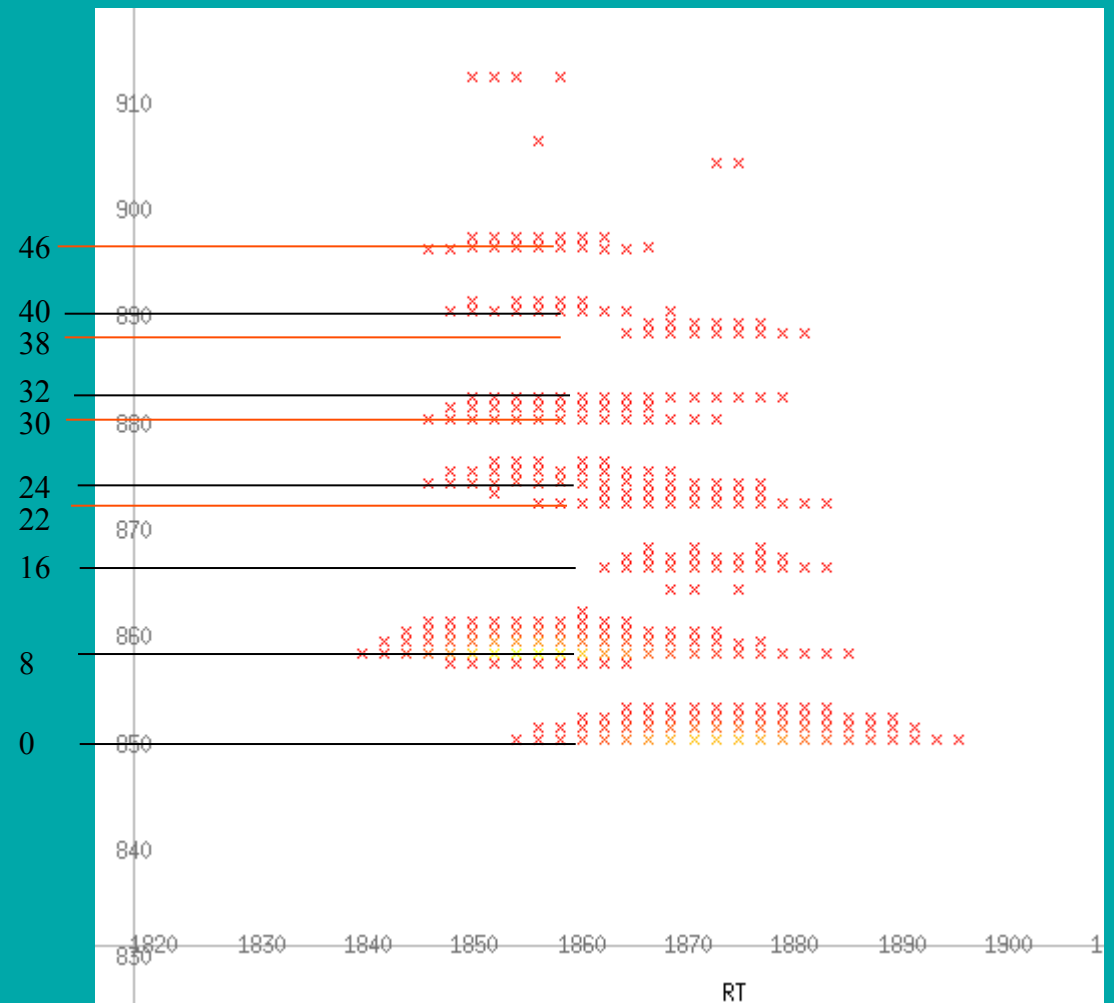
MA13\_02011\_02\_ALL01Z3I9A\* Overview (exhibits 'stack-ups')



CSE182

# Serum ICAT data

- Instead of pairs, we see entire clusters at 0, +8, +16, +22
- ICAT based strategies must clarify ambiguous pairing.



# ICAT problems

- Tag is bulky, and can break off.
- Cys is low abundance
- MS<sub>2</sub> analysis to identify the peptide is harder.

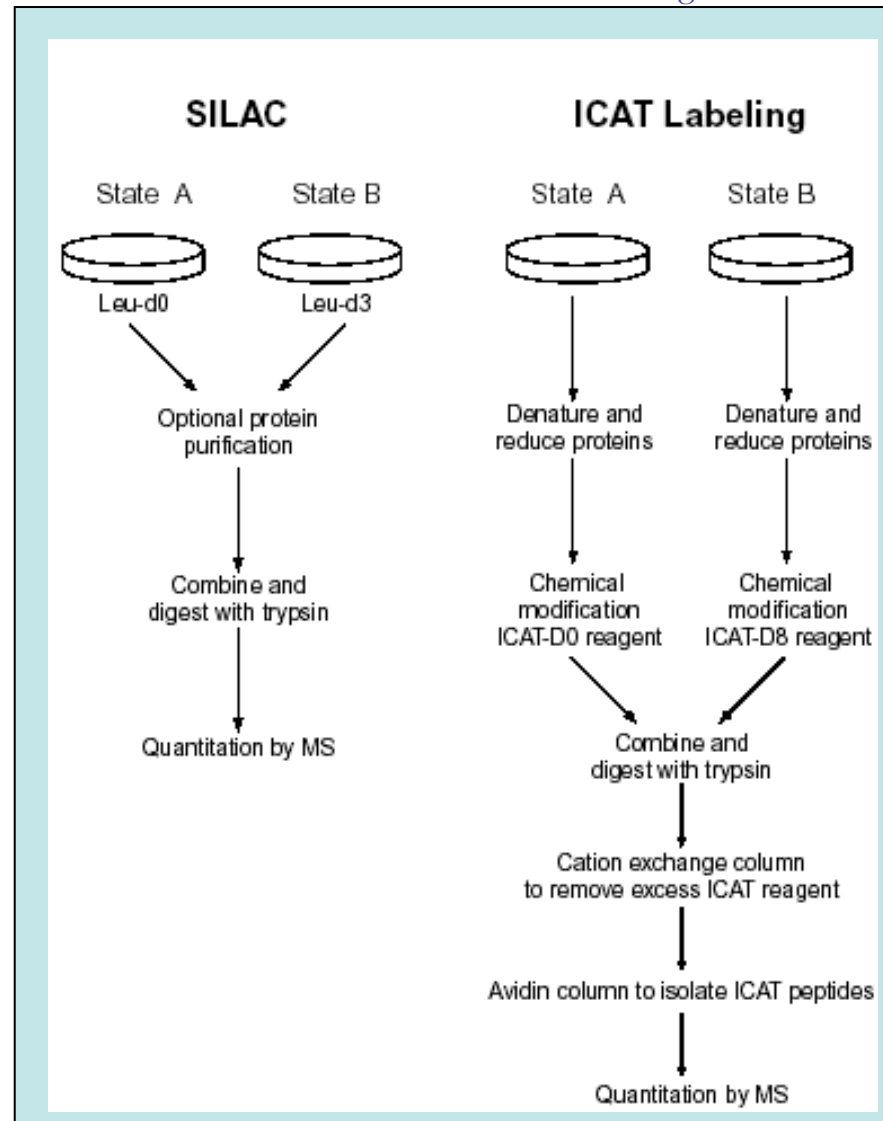
# SILAC

- A novel stable isotope labeling strategy
- Mammalian cell-lines do not 'manufacture' all amino-acids. Where do they come from?
- Labeled amino-acids are added to amino-acid deficient culture, and are incorporated into all proteins as they are synthesized
- No chemical labeling or affinity purification is performed.
- Leucine was used (10% abundance vs 2% for Cys)

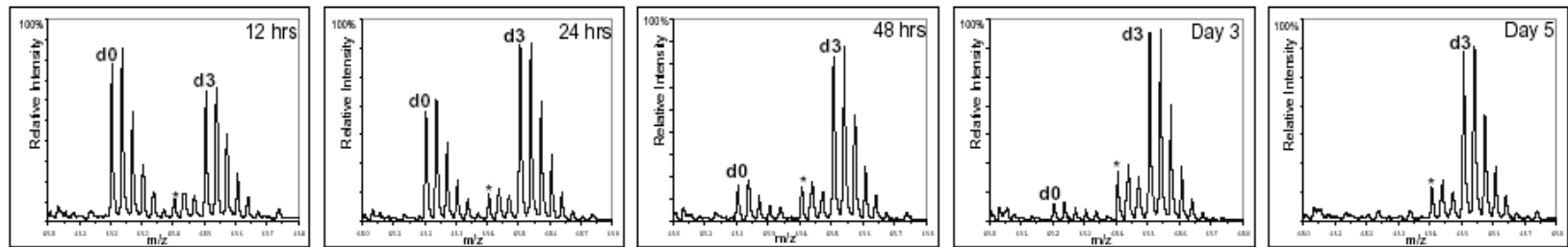
# SILAC vs ICAT

*Ong et al. MCP, 2002*

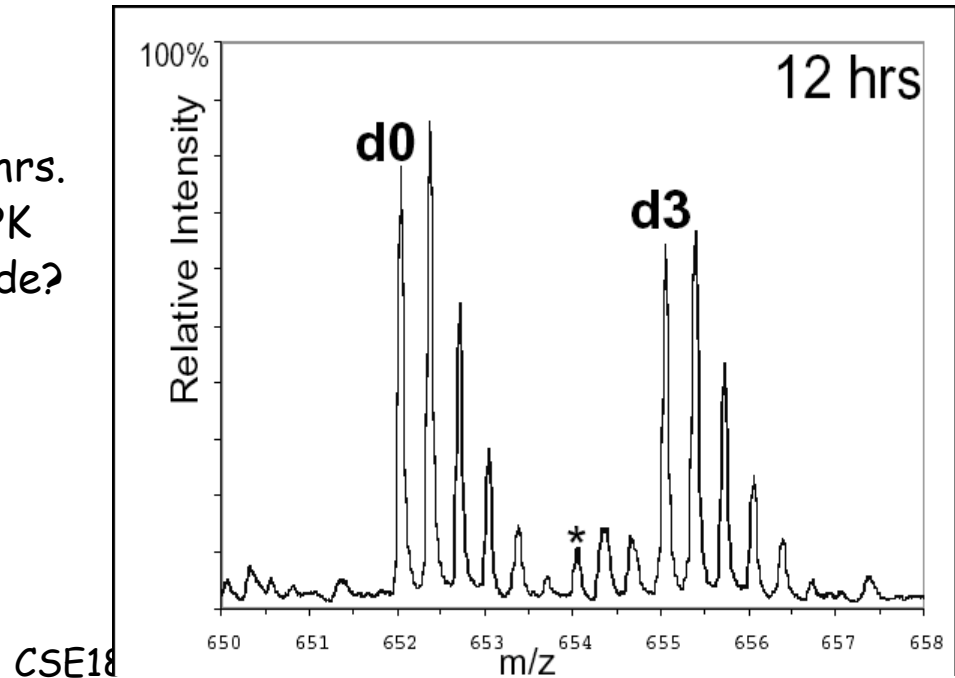
- Leucine is higher abundance than Cys
- No affinity tagging done
- Fragmentation patterns for the two peptides are identical
  - Identification is easier



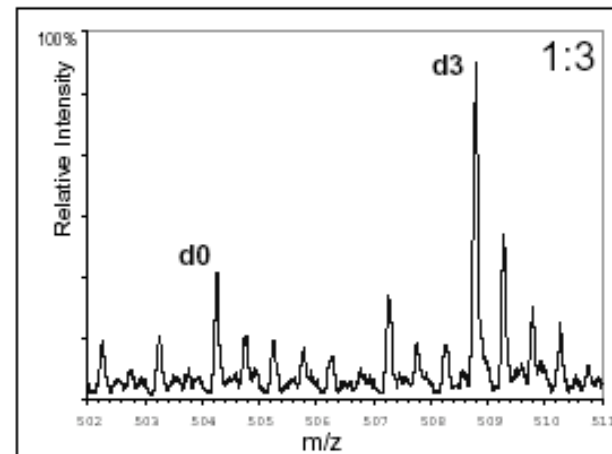
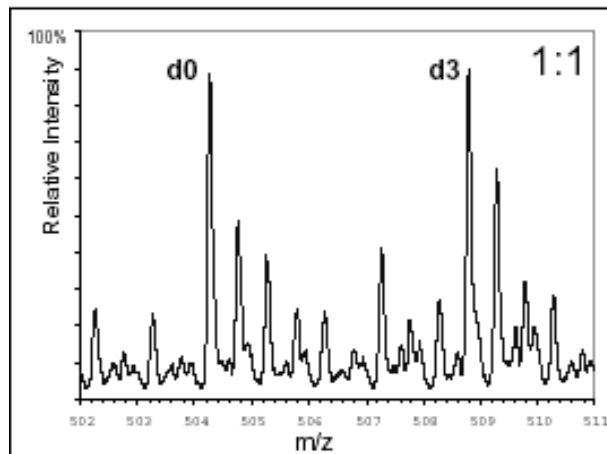
# Incorporation of Leu-d3 at various time points



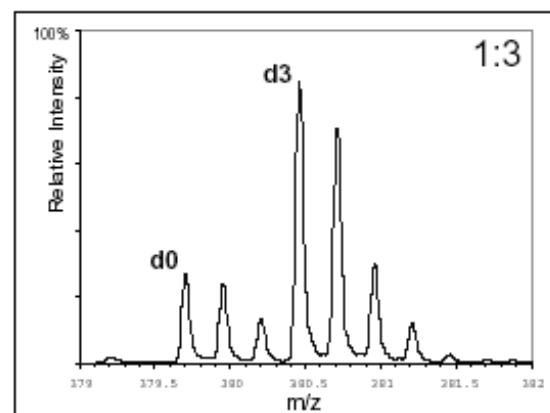
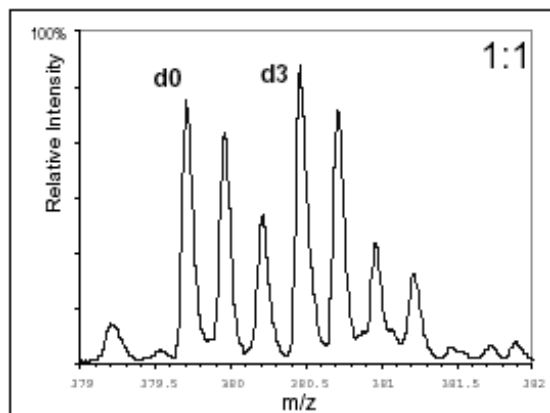
- Doubling time of the cells is 24 hrs.
- Peptide = VAPEEHPVLLTEAPLNPK
- What is the charge on the peptide?



# Quantitation on controlled mixtures



SCNCLLLK

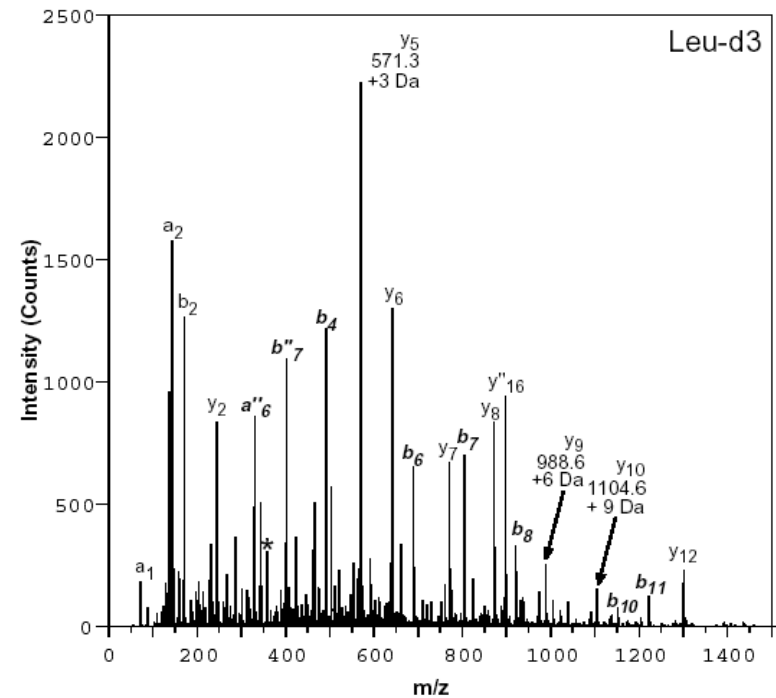
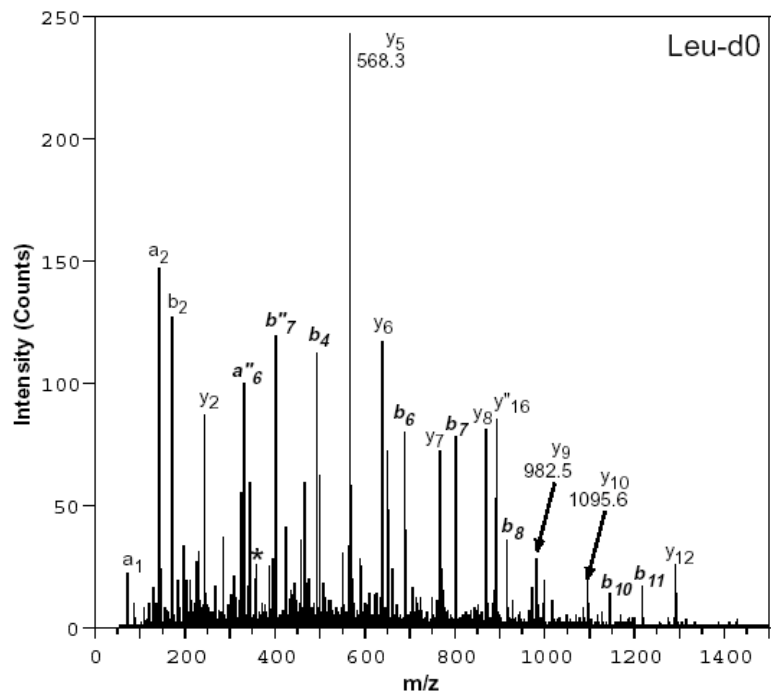


IWHHTFYNELR

CSE182

End of L13

# Identification



- MS/MS of differentially labeled peptides

# Peptide Matching

- SILAC/ICAT allow us to compare relative peptide abundances without identifying the peptides.
- Another way to do this is computational. Under identical Liquid Chromatography conditions, peptides will elute in the same order in two experiments.
  - These peptides can be paired computationally