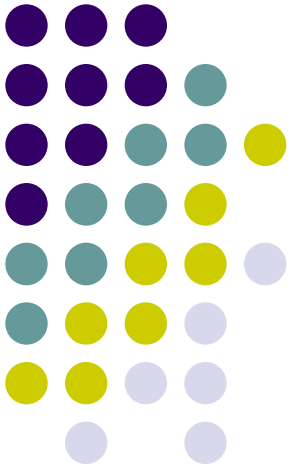
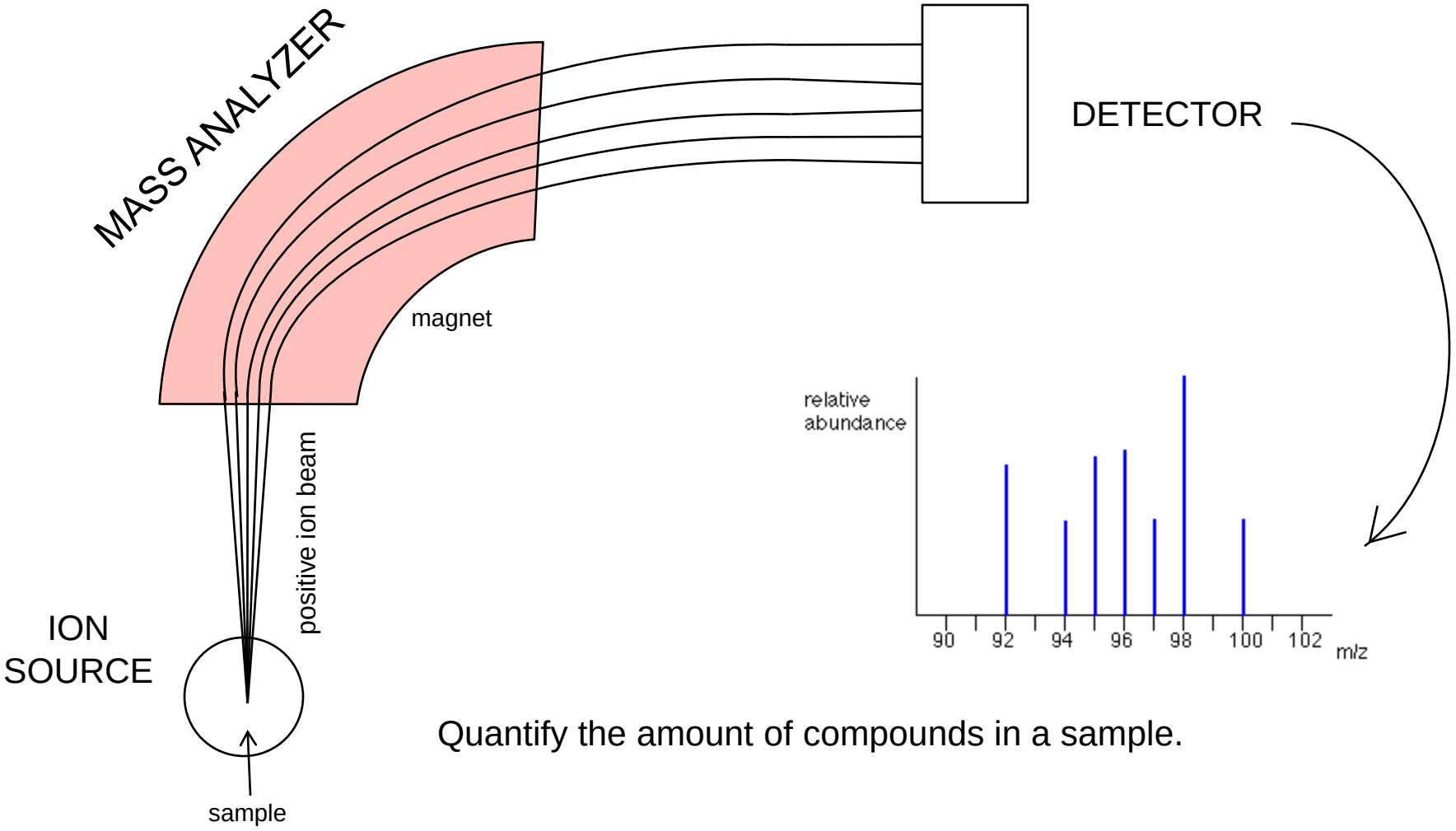
Shared Peptides in Mass Spectrometry Based Protein Quantification

Banu Dost, Nuno Bandeira,
Xiangqian Li, Zhouxin Shen, Steve Briggs,
Vineet Bafna

University of California, San Diego
contact: bdost@cs.ucsd.edu

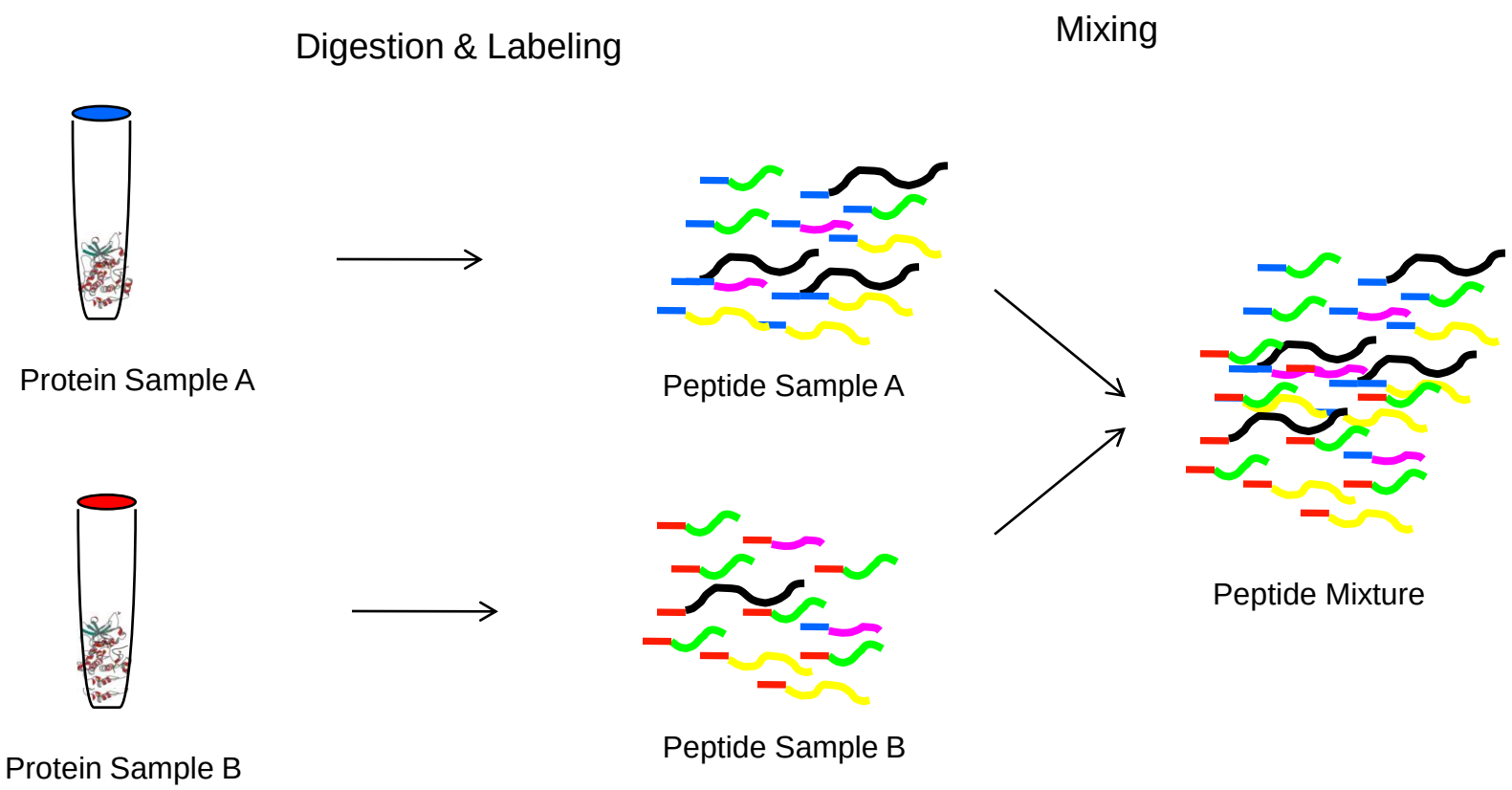


Mass Spectrometer

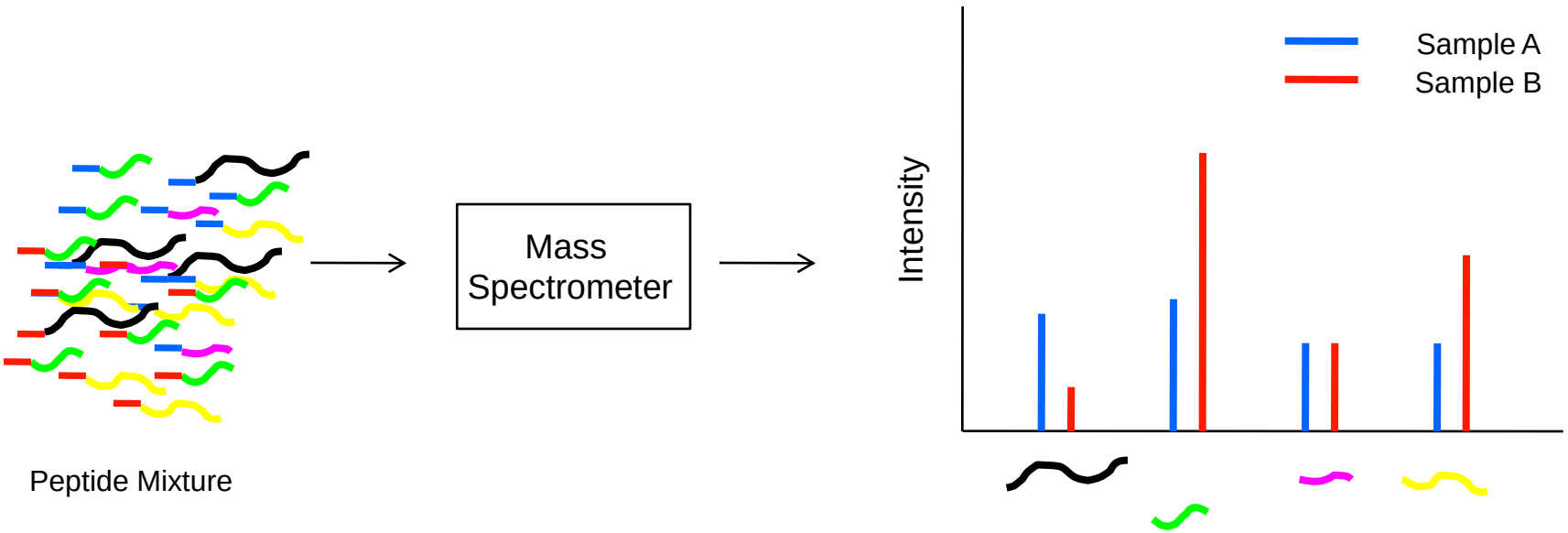


Quantify the amount of compounds in a sample.

Mass Spectrometry-based Protein Quantification

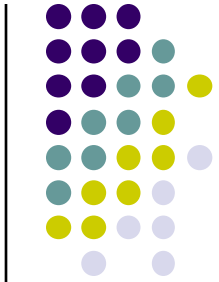


Mass Spectrometry-based Protein Quantification



Relative abundance of peptides across samples.

Mass Spectrometry-based Protein Quantification



- Traditionally,
  - If a protein does not have a unique peptide, its relative abundance across samples is not measured.
  - Relative abundance of 2 different proteins is never measured.

~50% of Peptides are shared

> ATPase 8

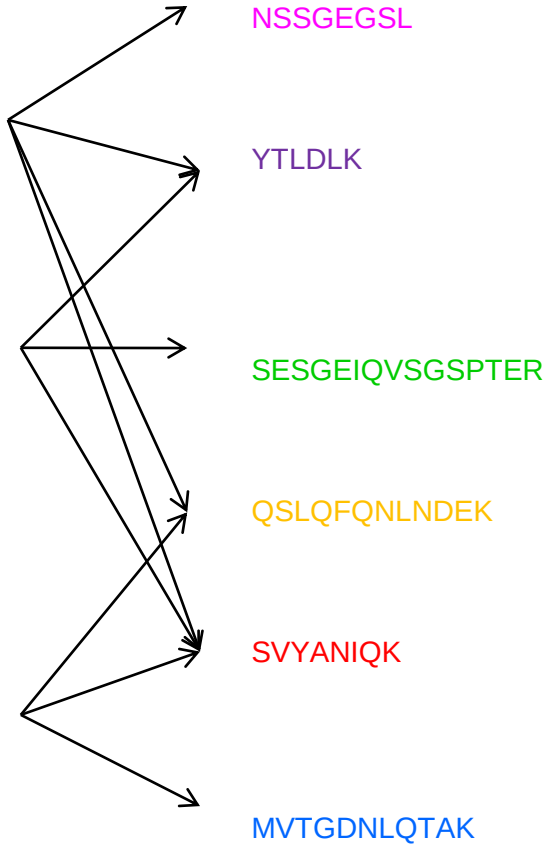
MTSLLKSSPGRRRGGDVESGKSEHADSDSDTFYIPSKNASIERLQQWRKAALVLN
ASRRFRYTLDLKKEQETREMRQKIRSHAHALLAANRFMDMGRESGVEK ...ILMVAA
VASLALGIKTEGIKEGWYDGGSIAFAVILVIVVTAVSDYKQSLQFQNLNDEKRNIHLE
…….NFASVVKVVRWGRSVYANIQKFIQFQLTVNVAALVINVVAAISSGDVPLTAVQ
LLWVNLIMDTLGALALATEPPTDHLMGRPPVGRKEPLITNIMWRNLLIQAIYQVSVLL
TLNFRGISILGLEHEVHEHATRVKNTIIFNAFVLCQAFNEFNARKPDEKNIFKGVIKNR
LFMGIIVITLVLQVIIVEFLGKFASTTKLNWKQWLICVGIGVISWPLALVGKFIPVPAAP
ISNKLKVLKFWGKKKNSSGEGSL

> ATPase 10

MSGQFNNSPRGEDKDVEAGTSSFTEYEDSPFDIASTKNAPVERLRRWRQAALVLN
ASRRFRYTLDLKREEDKKQMLRKMRAHAQAIRAA……GIA
HNTTGSVFRSESGEIQVSGSPTERAILNWAIKLG……..KSDIIILDDNFESVVKVVR
WGRSVYANIQKFIQFQLTVNVAALVINVVAAISAGEVPLTAVQLLWVNLIMDTLGALA
LATEPPTDHLMDRAPVGRREPLITNIMWRNLFIQAMYQVTVLLILNFRGISILHLKSKP
NAERVKNTVIFNAFVICQVFNEFNARKPDEINIFRGVLRNHLFVGIISITIVLQVVIVEFL
GTFASTTKLDWEMWLVCIGIGSISWPLAVIGKLIPVPETPVSQYFRINRWRRNSSG

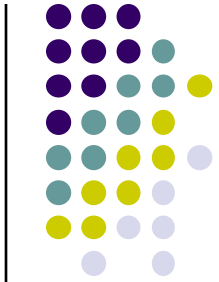
> ATPase 9

MSTSSSNGLLLTSMSGRHDDMEAGSAKTEEHSDHEELQHDPDDPFDIDNTKNASV
ESLRRWRQAALVLNASRRFRYTLDLNKEEHYDNRRRMIRAHAQVIRAALLFKLAGE
…….EKEVIDRKNAFGSNTYPKKKGKNFFMFLWEAWQDLTLIILIIAAVTSLALGIKT
EGLKEGWLDGGSIAFAVLLVIVVTAVSDYRQSLQFQNLNDEKRNIQLEV…..TLQSIE
SQKEFFRVAIDSMAKNSLRCVAIACRTQELNQVPKEQEDLDKWALPEDELILLAIVGI
KDPCRPGVREAVRICTSAGVKVRMVTGDNLQTAKAIALECGILSSDTEAVEPTIIEGK
VFRELSEKEREQVAKKITVMGRSSPNDKLLLVQALRKNGDVVAVTGDGTNDAPALH
EADIGLSMGISGTEVAKESSDIIILDDNFASVVKVVRWGRSVYANIQKFIQFQLTVNVA
ALIINVV……..GKLIPVPKTPMSVYFKKPFRKYKASRNA



[Based on arabidopsis ITRAQ data]

Shared Peptides



- ~50% of the peptides are shared by multiple proteins due to *[Jin et al., J. Proteome Res., 2008)]*
  - Homologues
  - Splicing variants (isoforms)

- ~50% of the proteins do not have a unique peptide.

- For protein quantification, only unique peptides are taken into consideration, half of the data is ignored.

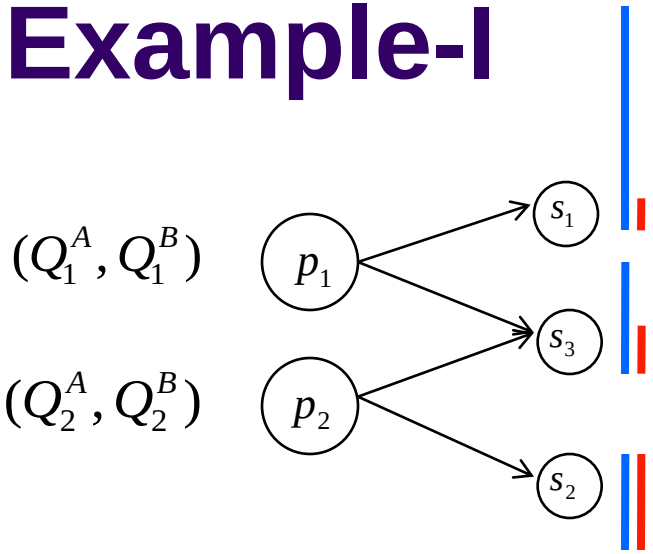
Goal

- Demonstrate that shared peptides are a resource that adds value to protein quantification.

1) Across-samples relative quantification of proteins with no unique peptide

2) Relative quantification of distinct proteins in a sample

Example-I



$r_1 = 16$ ➡ $Q_1^A = 16Q_1^B$

$r_3 = 4$ ➡ $(Q_1^A + Q_2^A) = 4(Q_1^B + Q_2^B)$

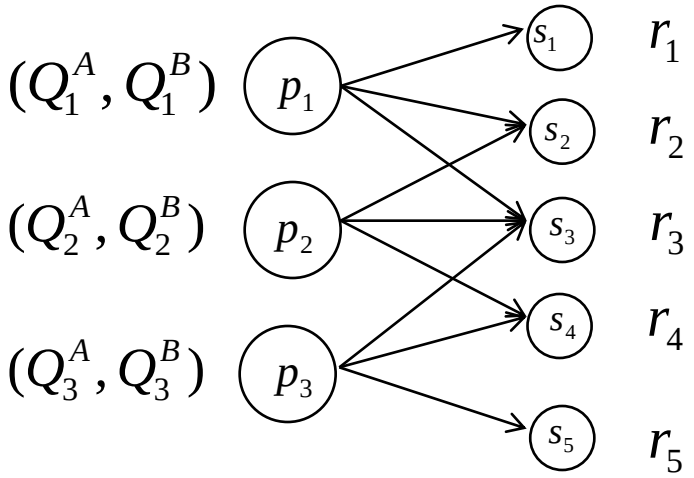
$r_2 = 1$ ➡ $Q_2^A = Q_2^B$

$$Q_1^A = 4Q_2^A$$

$$Q_2^B = 4Q_1^B$$

- We can compute relative abundances of proteins 1 & 2 within sample A & B.

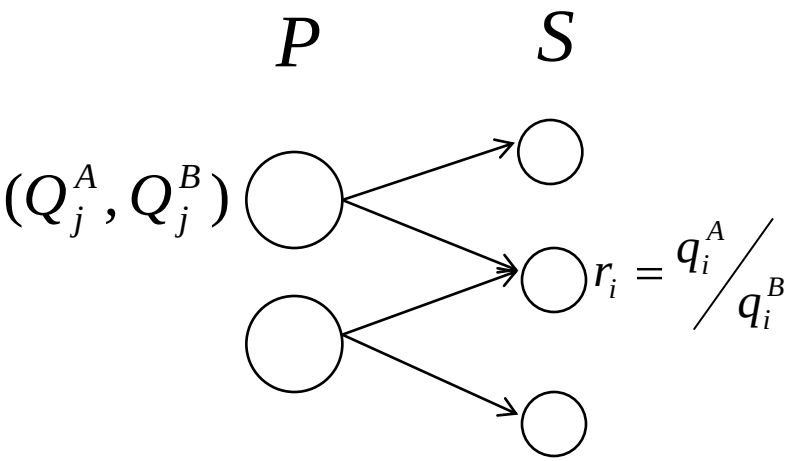
Example-II



6 unknowns,
5+1 constraints

- We can solve relative abundance of protein 2 across samples, even though it has no unique peptide.

Protein Quantification via Shared Peptides



P : set of proteins

S : set of peptides

Bi – partite graph $G = (P \cup S, E)$ where

$E = \{(p,s) | \, p \in P, s \in S, s \text{ is contained in } p\}$

Define variables denoting

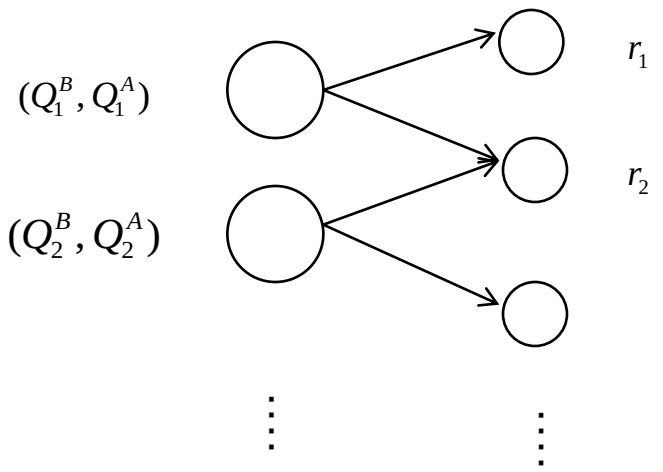
protein abundances $Q_j^A, Q_j^B \in R$ for all proteins.

s.t. $\sum Q_j^A = 100.$

For all peptide j, define $q_i^A, q_i^B, r_i = q_i^A / q_i^B \in R.$

Given r_j for all peptides, solve for $Q_i^A, Q_i^B \in R,$ for all proteins.

Linear Programming (LP) Problem Formulation



$$Q_1^A \approx r_1 Q_1^B$$

$$(Q_1^A + Q_2^A) \approx r_3 (Q_1^B + Q_2^B)$$

m=#proteins, n=#peptides
2m unknowns, n+1 constraints

Minimize $\varepsilon_1 + ... + \varepsilon_m$ subject to

$$\varepsilon_1 = |Q_1^A - r_1 \times Q_1^B|$$

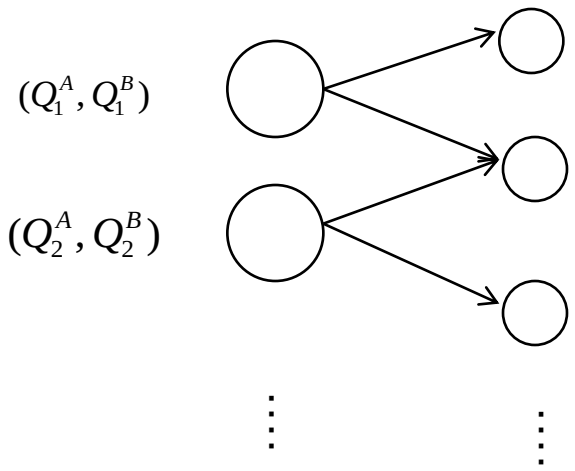
$$\varepsilon_2 = |(Q_1^A + Q_2^A) - r_2 \times (Q_1^B + Q_2^B)|$$

.....

$$\sum_{j=1:m} Q_m^A = 100$$

$$Q_j^A \geq 0, Q_j^B \geq 0$$

Linear Programming (LP) Problem Formulation



$$Q_1^A \approx r_1 Q_1^B$$

$$(Q_1^A + Q_2^A) \approx r_3 (Q_1^B + Q_2^B)$$

$$\min \left\| \vec{\boldsymbol{\varepsilon}} \right\|_1 \text{ where } \vec{\boldsymbol{\varepsilon}} = \mathbf{A}\vec{\mathbf{x}} - \vec{\mathbf{b}}, \vec{\mathbf{x}} > 0$$

$$\vec{\mathbf{x}} = [Q_1^A, \ldots Q_m^A, Q_1^B, \ldots, Q_m^B]^T$$

$$\vec{\mathbf{b}} = [100, 0, \ldots, 0]^T$$

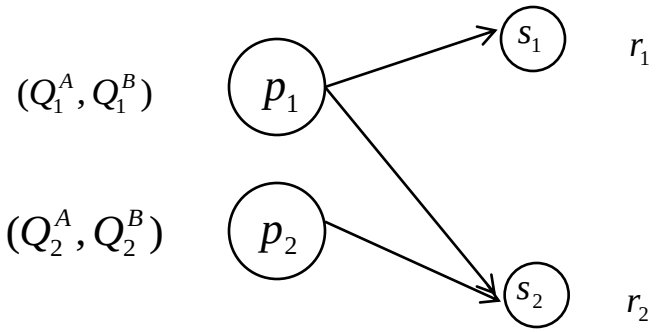
m :# proteins

n :# peptides

Given peptide ratios r_i,
we estimate *relative* protein amounts
Q^A_j and Q^B.

Robustness of Estimates

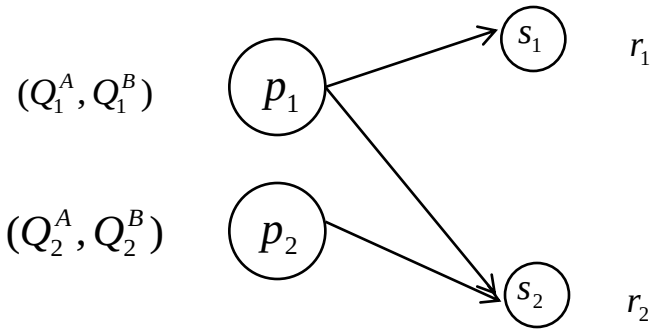
- Low objective does not necessarily result in robust estimates.



4 unknowns, 3 constraints
=> under-determined
=> infinitely many solutions with zero error

Robustness of Estimates

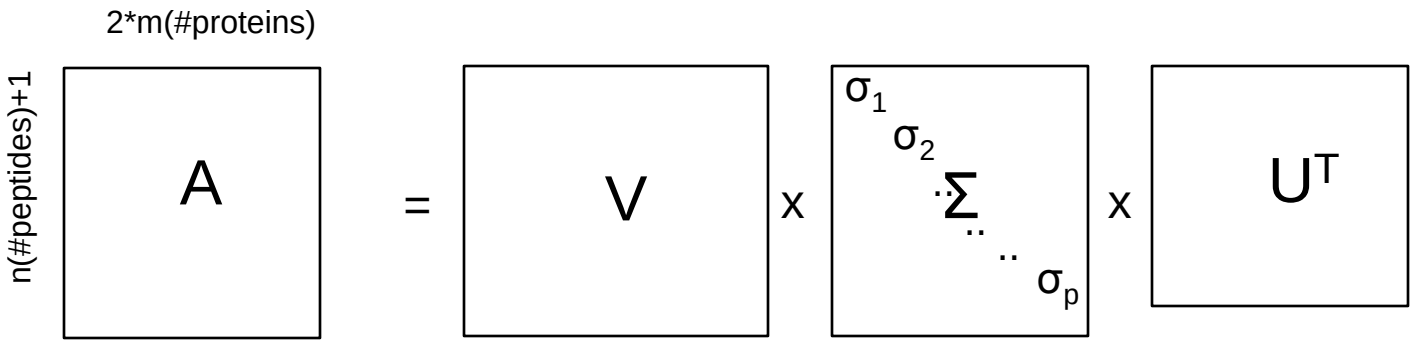
- Low objective does not necessarily result in robust estimates.



$$\min \left\| \vec{\boldsymbol{\varepsilon}} \right\|_1 \text{ where } \vec{\boldsymbol{\varepsilon}} = \mathbf{A}\vec{\mathbf{x}} - \vec{\mathbf{b}}, \vec{\mathbf{x}} > 0$$

Rank(A) < 2m => under-determined

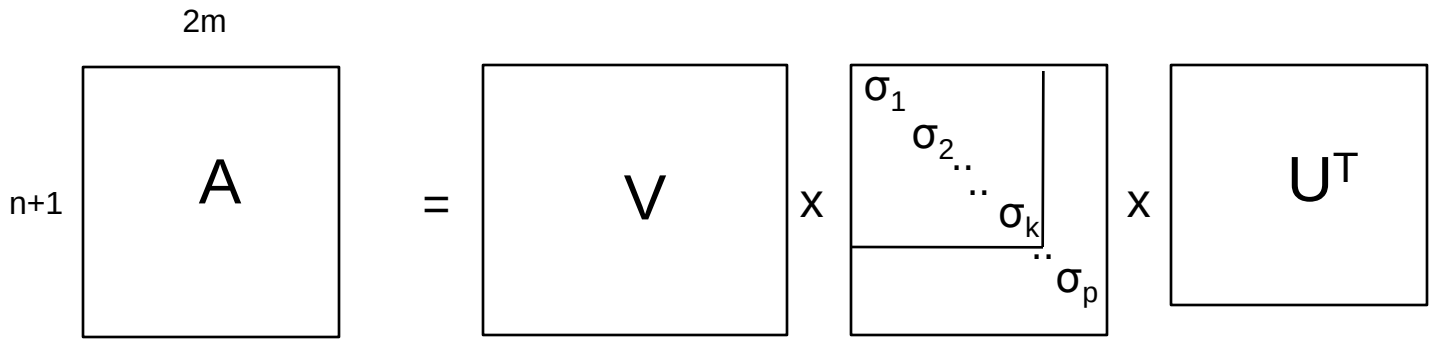
Rank-threshold



$$R(\mathbf{A}) = \min\{t \in Z^{+} \mid \sigma_j > 10^{-t}, \forall j\}$$

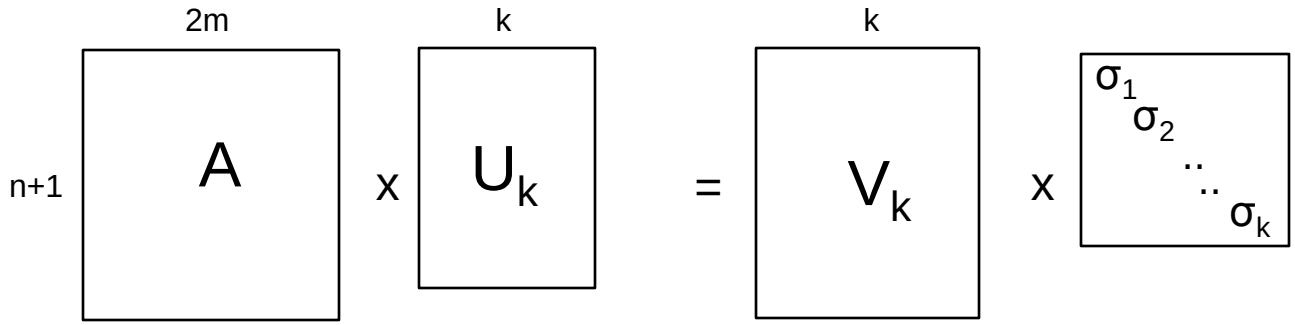
- A good way to characterize the reliability of the estimates.
  - R(A) = ∞ => under-determined system
  - R(A) is high => ∃ small singular values => ill-conditioned, poor estimates
  - R(A) is low => large singular values => full-rank, better estimates

Robust estimates for ill-conditioned systems



$$\text{if } k \text{ singular values} > 10^{-t}$$

Robust estimates for ill-conditioned systems



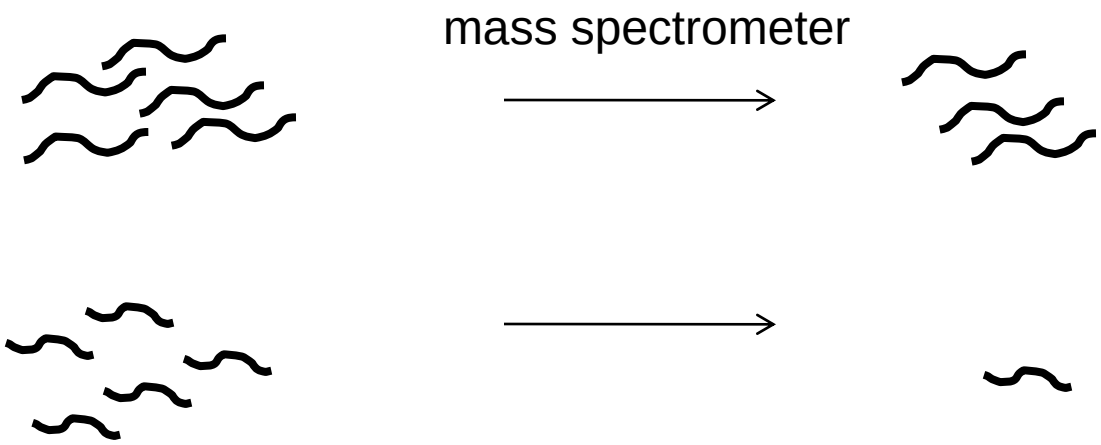
$$\text{if k singular values} > 10^{-t}, \text{then } R(\mathbf{AU_k}) = t$$

$$\min \left\| \vec{\boldsymbol{\varepsilon}} \right\|_1 \text{ where } \vec{\boldsymbol{\varepsilon}} = \mathbf{AU_k}\vec{\mathbf{y}} - \vec{\mathbf{b}}, \vec{\mathbf{y}} > 0, \mathbf{U_k y} > 0$$

$$\vec{\mathbf{x}} = \mathbf{U_k}\vec{\mathbf{y}}$$

Peptide Detectability

- Not all peptides are detected in mass spectrometer with the same efficiency.



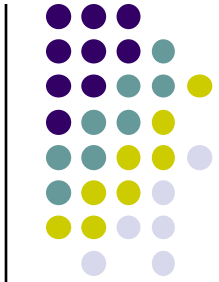
Incorporating Peptide Detectability

- Peptide detectability, $d_i \in [0,1]$
  - relates peptide abundance to the total abundances of its parent proteins.

$$q_i^A = d_i \times \sum_{(p_j, s_i \in E)} Q_j^A$$

$$q_i^B = d_i \times \sum_{(p_j, s_i \in E)} Q_j^B$$

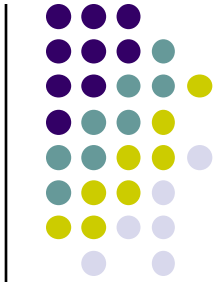
Incorporating Peptide Detectability



- If we know the detectabilities,
  - m=#proteins, n=#peptides
  - 2m variables, 2n+1 constraints (n more constraints)
  - The number of components solved are considerably increased.

- If we do not know the detectabilities,
  - 2m+n variables, 2n+1 constraints
  - Inference of peptide detectabilities in addition to relative protein abundances.

Incorporating Peptide Detectabilities

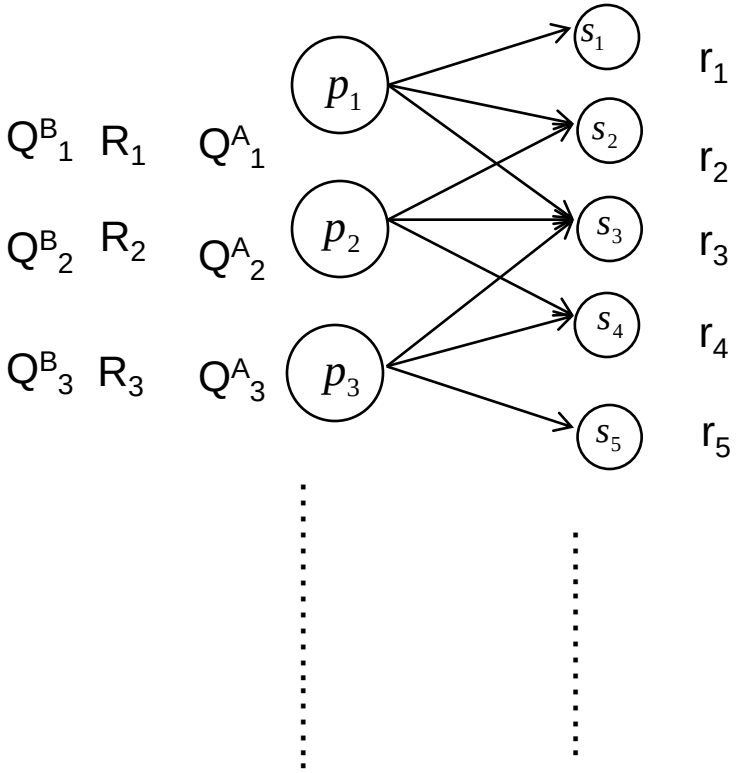


- Current mass spectrometry data do not provide reliable peptide abundance values.

- Recent developments indicate that
  - peptide abundances can be experimentally estimated. *[Bantscheff et al., Anal Bioanal Chem, 2007]*
  - peptide detectabilities can be reliably estimated across mass spectrometry runs. *[Alves, et al., Pac Symp Biocomput, 2007]*

Simulation

- Protein-peptide mapping based on Arabidopsis ITRAQ data.
  - 257 topologically different components
- Generate 100 datasets for each component.
  - $Q^B_j = R_j \times Q^A_j$
  - perturb according to a log-normal $N(0, \sigma)$.
    - σ : perturbation level
- Solve each dataset using LP formulation.



Simulation: Validation Statistics

If answer is known,

protein abundances distance

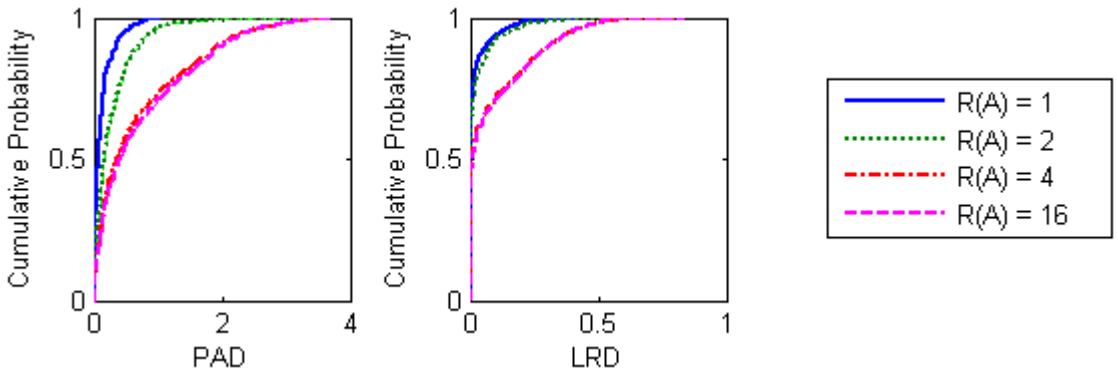
$$\mathrm{PAD}(\vec{\mathbf{Q}}^{A\prime}, \vec{\mathbf{Q}}^{A}) = \frac{\left\|\mathbf{R}^{A}\right\|_1}{m} \text{ where } \mathbf{R}^{A} = \left[\ln \frac{\vec{\mathbf{Q}}_i^{A\prime}}{\vec{\mathbf{Q}}_i^{A}}\right], m = \#\,\text{proteins}$$

If answer is unknown, we measure consistency by

peptide ratios distance

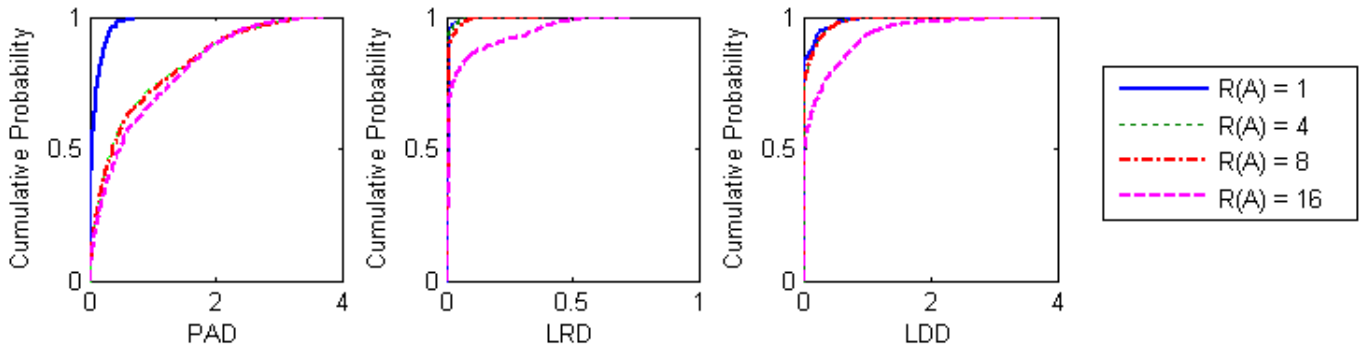
$$\mathrm{LRD}(\vec{\mathbf{r}}', \vec{\mathbf{r}}) = \frac{\left\|\vec{\mathbf{r}}' - \vec{\mathbf{r}}\right\|_1}{n} \text{ where } \vec{\mathbf{r}} = [\ln \mathbf{r}_i], \mathrm{n} = \#\,\text{peptides}$$

Simulation: Results



- With no noise, we achieve ideal case for all full-rank systems at R(A)=4.
- 1074 full-rank systems at R(A) = 1, σ=0.01.
  - 75% have PAD < 0.16 and LRD < 0.01.
- In all cases, objective is close to 0. ($<10^{-4}$)
- Performance degrades with less strict rank-thresholds.

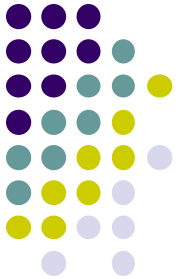
Simulation: Incorporating Peptide Detectabilities



- Peptide detectabilities distance

$$\mathrm{LDD}(\vec{d}', \vec{d}) = \frac{\left\| \log \vec{d}' - \log \vec{d} \right\|_1}{n}$$

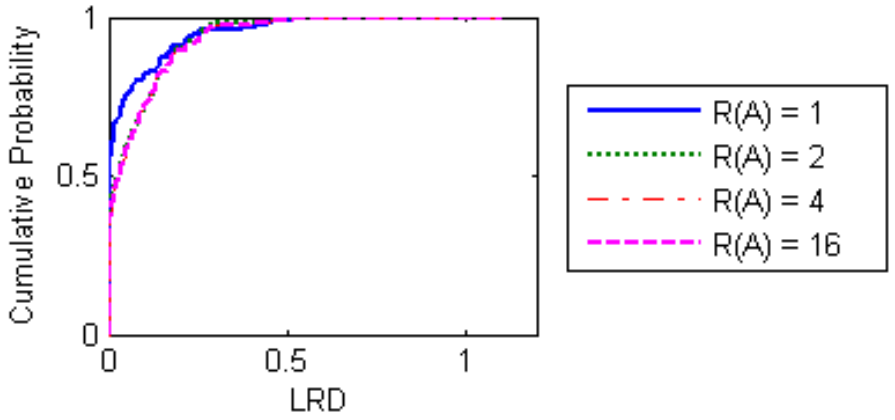
Arabidopsis ITRAQ Data



- Two samples before and after nematode infection
- ~120K spectra, 27K peptides mapping onto 8K protein
  - Close to half of the peptides (10K) are shared.
  - Close to half of the proteins (4K) do not have a unique peptide.
- Bi-partite mapping graph
  - 4119 connected components
  - 1190 have ≥2 proteins.
  - 257 non-isomorphic topologies, size ranging 2-127

Arabidopsis ITRAQ Data Results

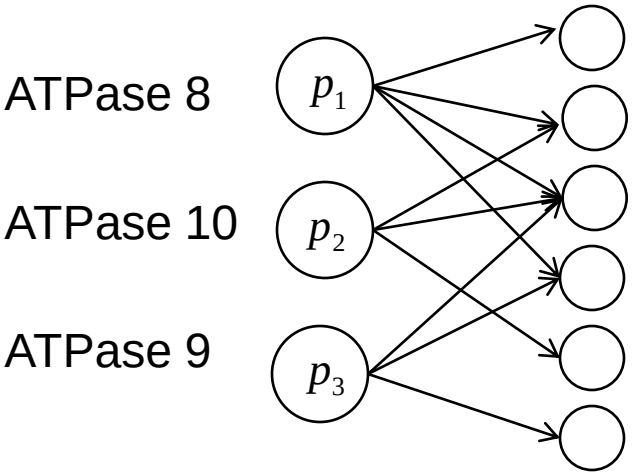
R(A)	#full-rank comps
1	99 (8.3%)
2	249 (20.9%)
4	276 (23.2%)
8	277 (23.3%)
16	282 (23.7%)



- 99 components with R(A)=1.
  - 219 proteins, 357 peptides
  - 79 have LRD < 10^{-1}, 55 have LRD < 10^{-4}

Arabidopsis ITRAQ Data Results – Example I

A system of 3 proteins from P-type Ca+2 ATPase super-family and 6 peptides.

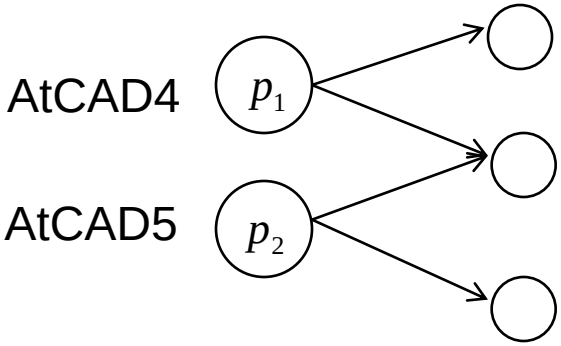


Q^A_1 : %34, R_1: 3.39, Q^B_1 : %66
Q^A_2 : %58, R_2: 0.95, Q^B_2 : %31
Q^A_3 : %8, R_3: 0.61, Q^B_3 : %3

ATPase8,10 are co-expressed evenly over all vegetative tissues [*Marmagne et al., Mol. Cell Proteomics, 2004*].

Arabidopsis ITRAQ Data Results – Example III

A system of 2 proteins from Cinnamyl-alcohol dehydrogeneases (CAD) family and 3 peptides.



Q^A_1 : %56, R_1: 1.5, Q^B_1 : %79
Q^A_2 : %44, R_2: 0.5, Q^B_2 : %21

Among many genes in CAD family members, only AtCAD4 and AtCAD5 are found to be central in the CAD metabolic network. [*Kim et al., Phytochemistry, 2007*]

Applications

- **Mass spectrometric data:**
  - Accurate peptide-protein mapping
  - Differential regulation of proteins from a family
  - Differential alternative splicing patterns
  - Differential phosphorylation

- **Transcript sequencing data:**
  - Accurate gene – exon mapping
  - Differential expression of transcripts

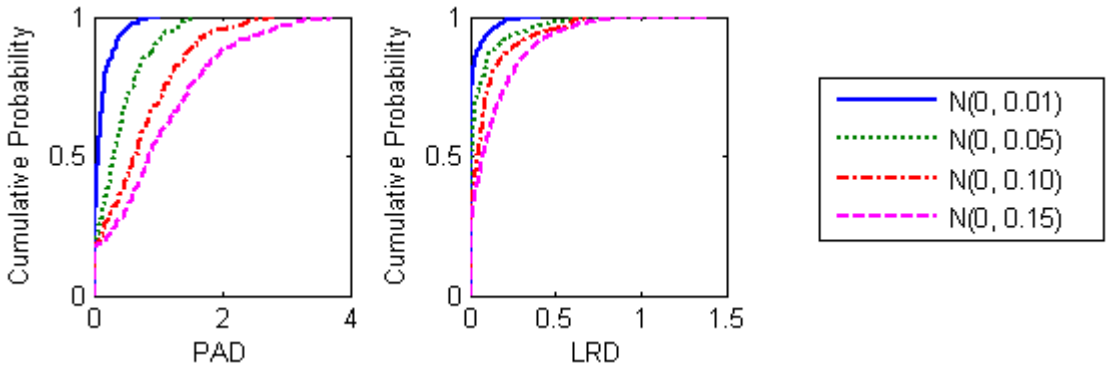
Discussion&Conclusion

- Shared peptides in protein quantification.
  - Relative abundance of proteins with no unique peptide
  - Relative abundance of distinct proteins.
- Accuracy of results depends upon the quality of the data.
- Viability of using shared peptides for peptide detectability computation

Acknowledgements

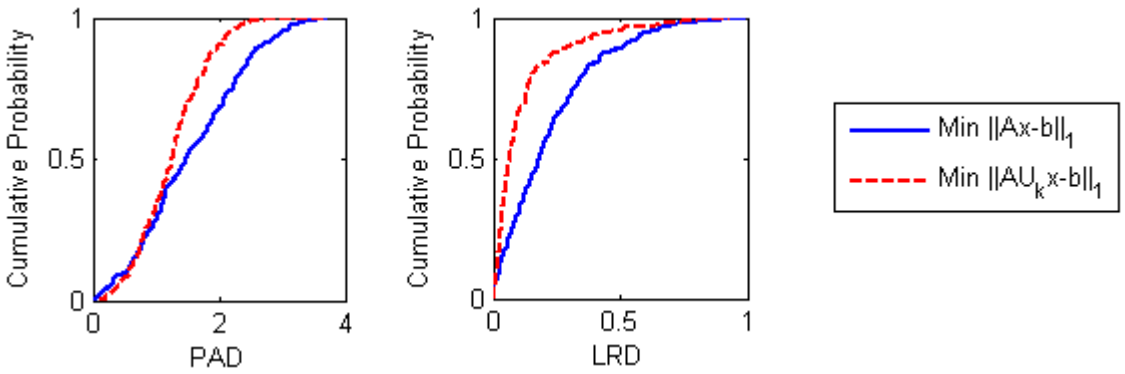
- Vineet Bafna (UCSD, CSE)
- Nuno Bandeira (UCSD, CSE)
- Xiangqian Li, Zhouxin Shen, Steve Briggs (UCSD, Biology)

Simulation: Results



- Performance degrades with an increase in perturbation error
- Full rank systems at R(A)=1, increasing perturbation levels
  - 93% have LRD ≤ 0.1 at σ=0.01
  - 55% have LRD ≤ 0.1 at σ=0.15

Result I: ill-conditioned systems



- 339 ill-conditioned systems
  - at most 3 singular values are $\leq 10^{-16}$, remaining are $\geq 10^{-1}$
- Revised LP for ill conditioned systems provides better estimates for those.
  - 65% have LRD ≤ 0.25 under original formulation
  - 88% have LRD ≤ 0.25 under revised formulation