UCSD Jacobs | Computer Science and Engineering

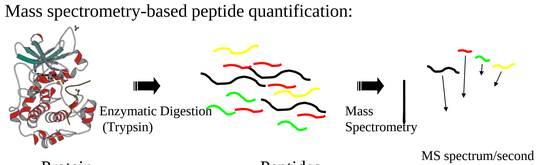
Center for Computational Mass Spectrometry

Shared Peptides in Mass Spectrometry Based Protein Quantification

Banu Dost, Nuno Bandeira, Vineet Bafna
Computer Science and Engineering Department, University of California, San Diego

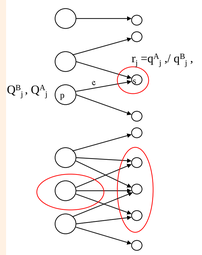
Introduction

Mass spectrometry-based peptide quantification:



- Peak-intensity for each peptide measured.
- Spectral intensity is a measurement of **abundance**, of peptide, i.e. the amount of its parent protein.

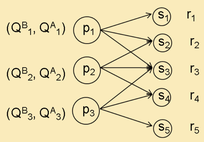
Protein-Peptide mapping graph:



- ~50% of the peptides are shared due to
  - 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.**

Aim

- **Protein Quantification via Shared Peptides:** Shared peptides provide extra information in protein quantification.



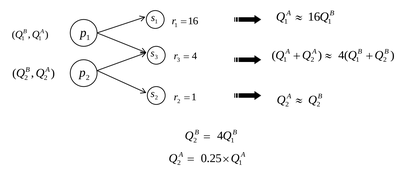
Our goal is to use shared peptides information to accurately compute

- across-samples relative abundance of each protein,
- relative abundances of distinct proteins in each sample.

given peptide abundance ratios across samples. (ri)

Method

Example:



$Q^B_2 = 4Q^B_1$

$Q^A_2 = 0.25 \times Q^A_1$

Linear Programming Problem Formulation: (F_1)

Input	Output	Formulation
		$\min \sum_{i=1}^{n} \lvert\varepsilon_i\rvert$
		s.t. $\sum_{p_j \in P} Q^B_j = 100$
$r_i, \forall s_i \in S$	$Q^B_j, Q^A_j, \forall p_j \in P$	$\varepsilon_i = \sum_{(p_j,s_i)\in E} Q^A_j - r_i \times \sum_{(p_j,s_i)\in E} Q^B_j \;\forall s_i \in S, r_i \geq 1$
		$\varepsilon_i = r_i \times \sum_{(p_j,s_i)\in E} Q^A_j - \sum_{(p_j,s_i)\in E} Q^B_j \;\forall s_i \in S, r_i \leq 0$
		$Q^B_j \geq 0, Q^A_j \geq 0 \qquad \forall p_j \in P.$

- Alternative formulation: $\min \sum_{i=1}^{n} \lvert\varepsilon_i\rvert$ where $\varepsilon = Ax - b, x > 0$
- Solvability of a system depends on the rank(A).
- rank(A) = #singular values > rank-threshold R(A).
  - Cat. I : Over-determined, full-rank systems
  - Cat. II : Ill-conditioned systems
  - Cat. III: Under-determined systems

Incorporating Peptide Detectabilities: (F_2)

- **Assumption:** one is able to estimate the absolute peptide abundances q^B_i and q^A_i.
- **Peptide detectability:** quantity $d_i \in [0, 1]$ that relates peptide abundance to the abundances of its parent proteins.

Input	Output	Formulation
		$\min \sum_{i=1}^{n} \lvert\varepsilon^B_i\rvert + \lvert\varepsilon^A_i\rvert$
		s.t. $\sum_{p_j \in P} Q^B_j = 100$
$q^B_i, q^A_i, r_i, \forall s_i \in S$	$Q^B_j, Q^A_j, \forall p_j \in P$	$\varepsilon^B_i = \sum_{(p_j,s_i)\in E} Q^B_j - q^B_i f_i, \forall s_i \in S$
	$d_i, \forall s_i \in S$	$\varepsilon^A_i = \sum_{(p_j,s_i)\in E} Q^A_j - q^A_i f_i, \forall s_i \in S$
		$Q^B_j \geq 0, \; Q^A_j \geq 0, \qquad \forall p_j \in P.$

- More robust
- More solvable components
- Peptide detectabilities inference

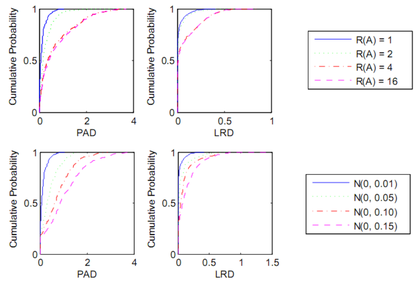
Results

Simulation:

- Based on Arabidopsis model with 258 non-isomorphic components
- 100 datasets for each.

	$\mathcal{R}(\mathcal{A})$				
	1	2	4	8	16
F_1 - Category I	1074 (4.2%)	2514 (9.8%)	3044 (11.8%)	3980 (15.5%)	4388 (17.1%)
F_2 - Category I	663 (2.6%)	2251 (8.8%)	2955 (11.5%)	3104 (12.1%)	4989 (19.4%)

- Validation statistics:
  - Protein abundance distance $\mathrm{PAD}(\vec{Q}^{B*}, \vec{Q}^B) = \frac{\lVert \vec{R}^B \rVert_1}{m}$ where $\vec{R}^B = \geq \left[\ln \frac{\vec{Q}^{B*}_i}{\vec{Q}^B_i} \right]_1$
  - Log peptide ratio distance $\mathrm{LRD}(\vec{r}^*, \vec{r}) = \frac{\lVert \ln \vec{r}^* - \ln \vec{r} \rVert_1}{n}$



Cdf of PAD and LRD for Category I systems (a) perturbation level 0.01, but different rank-thresholds (b) at rank-threshold 1, but different perturbation levels.

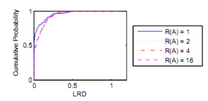
Arabidopsis ITRAQ Data:

- 1190 components, covering 8K proteins, 27K peptides.

(a)

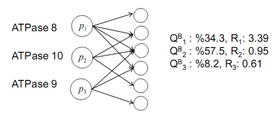
$\mathcal{R}(\mathcal{A})$	Cat. I	Cat. II
1	99 (8.3%)	191 (16%)
2	249 (20.9%)	41 (3.4%)
4	276 (23.2%)	14 (1.2%)
8	277 (23.3%)	13 (1.1%)
16	282 (23.7%)	8 (0.7%)

(b)

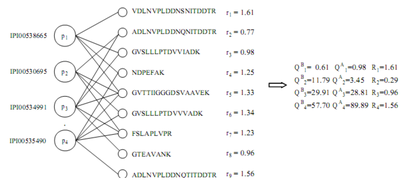


(a) Number of systems at different rank-thresholds.
(b) (b) ecdf of LRD for Category I systems at different rank-thresholds.

Example I: P-type Ca+2 ATPase super-family



Example II: PhosphoGlycerate kinase family



Applications

- Approach can easily be modified to apply on different datasets.

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

- Our work the **first attempt** to use shared peptides in **protein quantification**.
- Using shared peptides, we recover relative abundance values of distinct proteins in each sample and across-sample abundance ratio of each protein.
- Our results attest to the viability of using shared peptides for **detectability computation**.
- We investigate topological and numerical considerations in estimating **reliability of our computations.**
- Final quality of the results depends upon the accuracy of the experimental abundance computations. As the mass spectrometers become more accurate, the power of our methods will increase.
- Different algorithms can be used to optimize the error in estimation, including **non-linear optimization** and other **machine learning** approaches.
- We have experimented using s**imulated annealing** approach with a non-linear cost function.

Acknowledgements

The research was supported by the National Center for Research Resources of NIH via grant P-41-RR24851.