Proteomics Seminar 2014SS

Xiao Liang

April 29, 2014

BIOINFORMATICS ORIGINAL PAPER

Vol. 28 no. 22 2012, pages 2956–2962
doi:10.1093/bioinformatics/bts540

Data and text mining

Advance Access publication September 6, 2012

A linear programming model for protein inference problem in shotgun proteomics

Ting Huang and Zengyou He*

School of Software, Dalian University of Technology, Dalian 116621, China

Associate Editor: Jonathan Wren

Part I

Protein Inference Problem

Protein inference in shotgun proteomics experiment

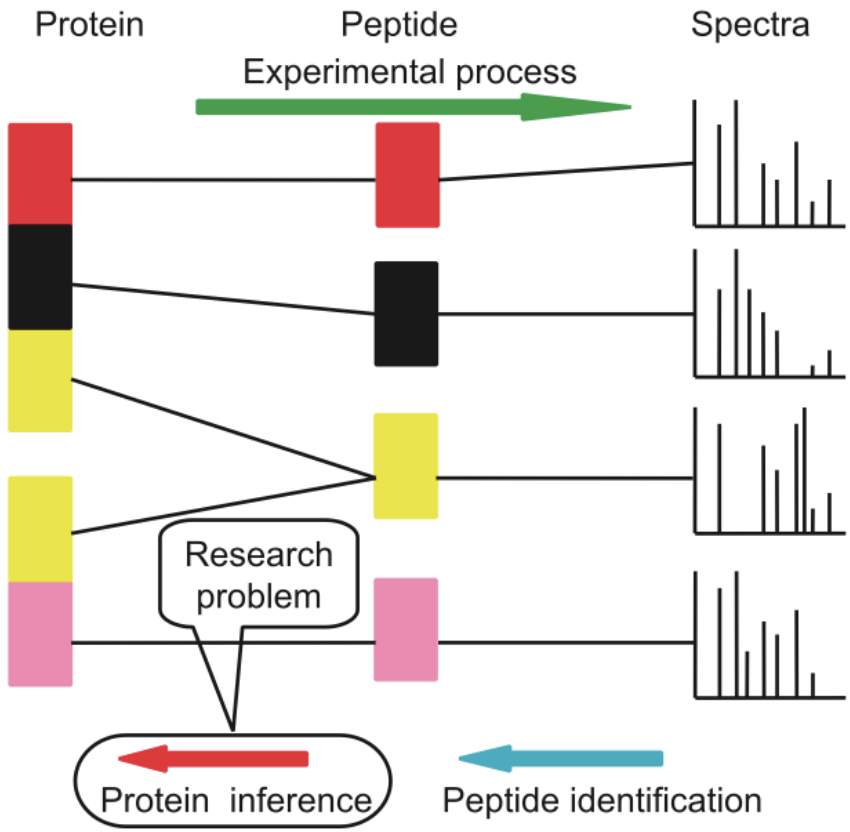
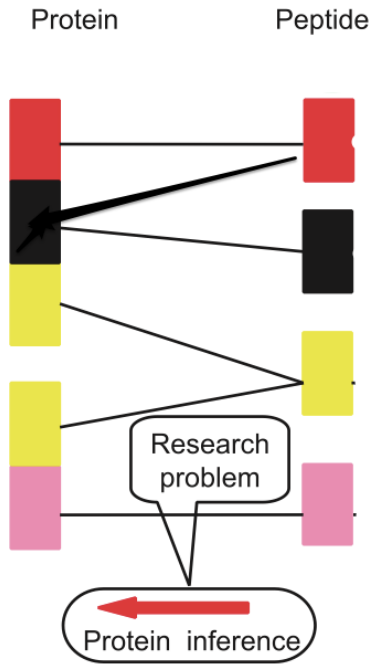


Figure: Protein inference using mass spectromery data.[1]

Goal: Find a subset of proteins that are truly present in the sample.

[1] Ting Huang and Zengyou He. "A linear programming model for protein inference problem in shotgun proteomics." In: *Bioinformatics* 28.22 (2012), pp. 2956–2962.

What input do we need for protein inference?



- A list of identified peptides.
  1. Database-driven approach
  2. de novo algorithm
- Peptide probabilities (detecbilities). <- rigorous statitical validation
  PeptideProphet[2] estimates $Pr(+|S)$: the probability that the peptide assignment with discriminant score S is correct.
- A list of candidate proteins.
- Expected output: a set of proteins accompanying protein probabilities.

[2] A. Keller et al. "Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search". In: *Analytical chemistry* 74.20 (2002), pp. 5383–5392.

Challenge: Peptide degeneracy

Peptide degeneracy: a single peptide mapped to multiple proteins.

Peptides	Proteins	Peptide probabilities
ASTSSSSSSSSNQQTEKETNTPK	P51965|UB2E1_HUMAN	0.9970
YEWRSTILGPPGSVY	P51965|UB2E1_HUMAN	0.6467
YEWRSTILGPPGSVY	Q96LR5|UB2E2_HUMAN	0.6467
YEWRSTILGPPGSVY	Q969T4|UB2E3_HUMAN	0.6467
VLLSICSLLTDCNPADPLVGSIATQYMTNR	P51965|UB2E1_HUMAN	1.0000

Figure: Peptide identifications (Sigma49 data)

- Shared peptides should belong to all proteins that they can match.

Challenge: Peptide degeneracy

Peptide degeneracy: a single peptide mapped to multiple proteins.

Peptides	Proteins	Peptide probabilities
ASTSSSSSSSSNQQTEKETNTPK	P51965|UB2E1_HUMAN	0.9970
YEWRSTILGPPGSVY	P51965|UB2E1_HUMAN	0.6467
YEWRSTILGPPGSVY	Q96LR5|UB2E2_HUMAN	0.6467
YEWRSTILGPPGSVY	Q969T4|UB2E3_HUMAN	0.6467
VLLSICSLLTDCNPADPLVGSIATQYMTNR	P51965|UB2E1_HUMAN	1.0000

Figure: Peptide identifications (Sigma49 data)

- Shared peptides should belong to all proteins that they can match.
- Conditional probability: model the conditional probability of
  - one protein being present given a peptide,
  - one peptide being present given a protein.

Existing protein inference algorithms

- ProteinProphet[3] calculates the conditional probability.
  Given peptides $i, i = 1 \cdots n$, with probabilities $Pr(+|S_i)$ corresponding to a protein , the probability p that this protein is present:

$$p = 1 - \prod_{i}^{n} [1 - Pr(+|S_i)]. \tag{1}$$

- Fido[4] estimates the protein posterior error probability.

$$p = Pr(+|protein). \tag{2}$$

[3] Alexey I Nesvizhskii et al. "A statistical model for identifying proteins by tandem mass spectrometry". In: *Analytical chemistry* 75.17 (2003), pp. 4646–4658.

[4] Oliver Serang, Michael J MacCoss, and William Stafford Noble. "Efficient marginalization to compute protein posterior probabilities from shotgun mass spectrometry data". In: *Journal of proteome research* 9.10 (2010), pp. 5346–5357.

Part II

Protein linear programming (ProteinLP)

Model

Table 1. Notations and definitions

Notations	Definitions
$(1, \ldots, i, \ldots, n)$	All n peptides identified by peptide identification algorithms
$(1, \ldots, j, \ldots, m)$	All m proteins that might have generated these n peptides
$(y_1, \ldots, y_i, \ldots, y_n)$	Peptide vector: indicator variables of peptides' presences if peptide i is present, $y_i = 1$; otherwise $y_i = 0$
$(x_1, \ldots, x_j, \ldots, x_m)$	Protein vector: indicator variables of proteins' presences
$(z_1, \ldots, z_i, \ldots, z_n)$	The probabilities of peptides' presences estimated by peptide identification algorithms or PeptideProphet

- $Pr(x_j = 1)$: the probability that protein j is present in the sample.
- $Pr(y_i = 1, x_j = 1)$: the probability that peptide i and protein j are present in the sample.

$$Pr(x_j = 1) = 1 - \prod_{i=1}^{n} [1 - Pr(y_i = 1, x_j = 1)] \tag{3}$$

Model

- From Eq. 3:

$$Pr(x_j = 1) = 1 - \prod_{i=1}^{n}[1 - Pr(y_i = 1, x_j = 1)] = 1 - \prod_{i=1}^{n} e^{\ln[1 - Pr(y_i = 1, x_j = 1)]}. \quad (4)$$

The protein probability is rewritten as:

$$Pr(x_j = 1) = 1 - \prod_{i=1}^{n} e^{p_{ij}}, \quad (5)$$

where $p_{ij} := \ln[1 - Pr(y_i = 1, x_j = 1)] \leq 0$.

- The peptide probability:

$$Pr(y_i = 1) = 1 - \prod_{j=1}^{m}[1 - Pr(y_i = 1, x_j = 1)] = 1 - \prod_{j=1}^{m} e^{p_{ij}}. \quad (6)$$

$$z_i = 1 - \prod_{j=1}^{m} e^{p_{ij}} \quad (7)$$

LP formulation

Objective:

Maximize the number of proteins with zero probabilities,

while peptide probabilities from joint probabilities should be as close to the input value as possible.

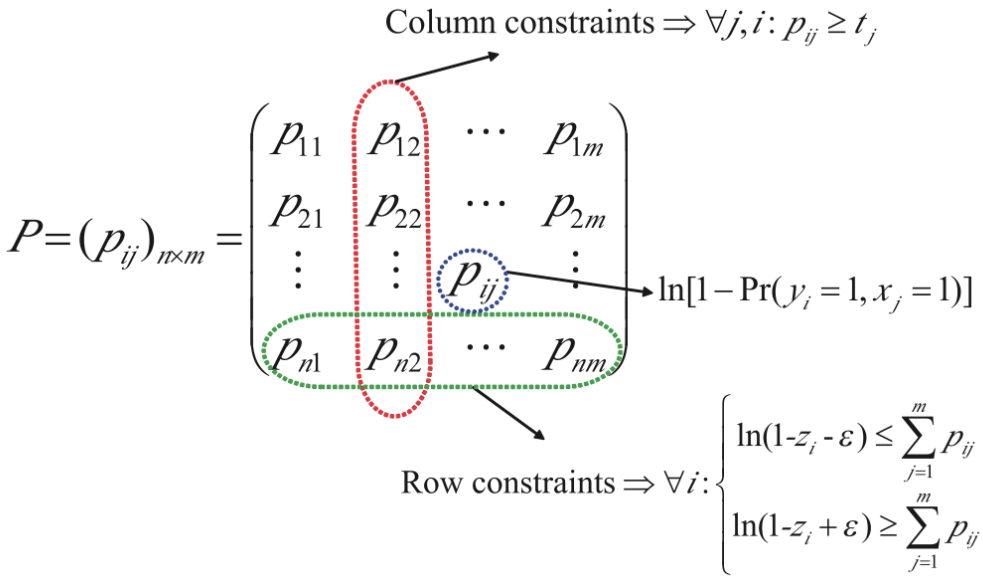
$$\text{Maximize:} \quad \sum_{j=1}^{m} t_j, \tag{8}$$

$$\text{Subject to:} \quad \forall i : t_j \leq p_{ij} \leq 0, \tag{9}$$

$$\forall i : \ln(1 - z_i - \epsilon) \leq \sum_{j=1}^{m} p_{ij} \leq \ln(1 - z_i + \epsilon), \tag{10}$$

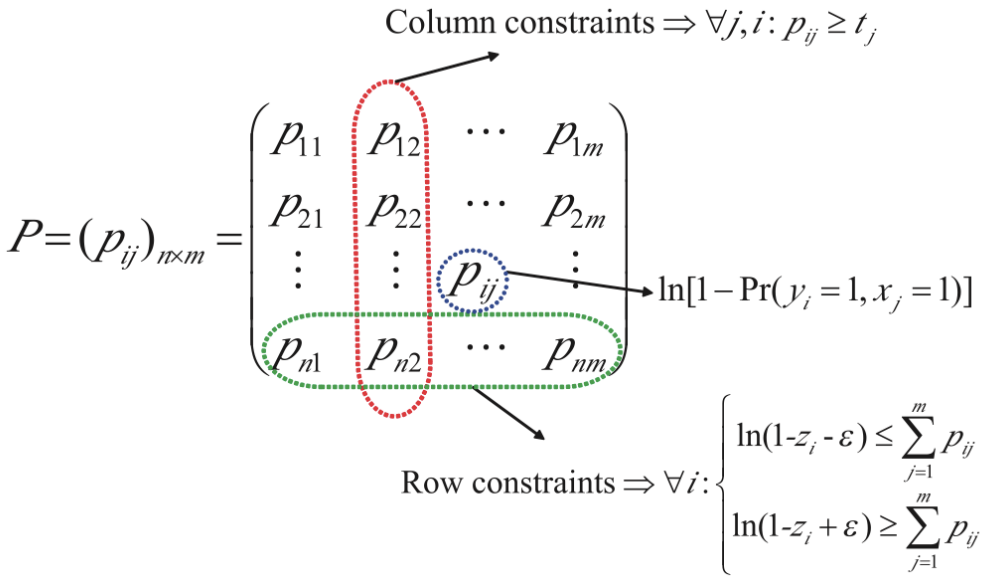
$$p_{ij} = 0, \quad \text{if protein j doesn't contain peptide i.} \tag{11}$$

LP formulation



- Constraint (11):
  - $p_{ij} = 0$ if $Pr(y_i = 1, x_j = 1) = 0$.

LP formulation



- Constraint (11):
  - $p_{ij} = 0$ if $Pr(y_i = 1, x_j = 1) = 0$.
- Constraint (10) peptide probability:

$$z_i \pm \epsilon = 1 - \prod_{j=1}^{m} e^{p_{ij}} \quad (12)$$

$$\Rightarrow \ln(1 - z_i \pm \epsilon) = \sum_{j=1}^{m} p_{ij} \quad (13)$$

LP formulation solved with GLPK

A standard LP:

$$\begin{aligned} \text{Maximize:} \quad & c^T x + c_0, && (14)\\ \text{Subject to:} \quad & Ax = b, && (15)\\ & Ax \leq b, && (16)\\ & LB \leq x \leq UB. && (17) \end{aligned}$$

$$x = (p_{11} \cdots p_{1m}\ p_{21} \cdots p_{2m} \cdots p_{nm}\ t_1 \cdots t_m)^T$$

$$c^T = (0|\mathbf{1}_{1,m}), \quad c_0 = 0$$

$$A = \left(\begin{array}{cccc|c} \mathbf{1}_{1,m} & \mathbf{0} & \cdots & \mathbf{0} & \\ \mathbf{0} & \mathbf{1}_{1,m} & & & \\ & & \vdots & & \mathbf{0} \\ \mathbf{0} & & & \mathbf{1}_{1,m} & \end{array}\right)$$

LP formulation solved with GLPK

Using results of P from Glpk, joint probability matrix $1 - e^P$ is computed :

	$Protein1$	$Protein2$	$Protein3$	$\cdots$	$Protein_m$
Peptide 1 (0.9)	0.9	0	0	$\cdots$	0
Peptide 2 (0.85)	0.7	0.5	0	$\cdots$	0
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	
Peptide n (0.9)	0	0.5	0	$\cdots$	0.8
Protein Probabilities	$1 - (1 - 0.9)(1 - 0.7)$	0.75	0	$\cdots$	0.8

Peptide degeneracy

- ProteinLP: joint probability $Pr(x_j = 1, y_i = 1)$. e.g. if peptide i present in more than one protein: m, n, r:

$$Pr(x_m = 1, y_i = 1) \cdot Pr(x_n = 1, y_i = 1) \cdot Pr(x_r = 1, y_i = 1) > 0 \tag{18}$$

- ProteinProphet: taking a weight w_i^n into account, if peptide i corresponds to N different proteins.

$$p_n = 1 - \prod_i^n (1 - w_i^j Pr(+|S_i)) \quad j = 1 \cdots N. \tag{19}$$

Combining with Number of Sibling Peptides (NSP): $NSP_i = \sum_{\{m|m \neq i\}} p(+|D_m)$.

$$Pr(+|S, NSP) = \frac{Pr(S|+)Pr(NSP|+)}{Pr(S|+)Pr(NSP|+) + Pr(S|-)Pr(NSP|-)}. \tag{20}$$

Part III

Results

Datasets

- Ground-truth data: 18 mixtures (Klimek et al., 2008), Sigma49 and yeast (Ramakrishnan et al., 2009a)

- Data without reference sets: DME (Brunner et al., 2007), HumanMD (Ramakrishnan et al., 2009b) and HumanEKC (Ramakrishnan et al., 2009a).

Sigma49 tested

- Data obtained from `http://www.marcottelab.org/MSdata/` .
- Peptide identification: X!Tandem (v2010.10.01.1) (David and Cottrell, 2004).
- GLPK (LPWrapper in OpenMS)
- Proteinlists:

```
</ProteinIdentification>
<PeptideIdentification score_type="XTandem" higher_score_better="true" significance_threshold="0" MZ="667.96337890625" RT="901.678" >
    <PeptideHit score="20.8" sequence="DQQKDAEGEGLSATTLLPK" charge="3" aa_before="K" aa_after="L" protein_refs="PH_4025" >
        <UserParam type="float" name="E-Value" value="1.1"/>
    </PeptideHit>
</PeptideIdentification>
<PeptideIdentification score_type="XTandem" higher_score_better="true" significance_threshold="0" MZ="408.515991210938" RT="902.077" >
    <PeptideHit score="23.2" sequence="SPPPSPTTQRR" charge="3" aa_before="R" aa_after="L" protein_refs="PH_432 PH_429 PH_428 PH_4628 PH_5036" >
        <UserParam type="float" name="E-Value" value="0.65"/>
    </PeptideHit>
</PeptideIdentification>
```

Validation

With setting a threshold t on the protein probabilities, only positive proteins remain.

False positives can be determined:

- Ground truth datasets.
- Datasets without references - using target-Decoy Analysis.
  - Protein database comtaminated with a set of shuffled unreal sequences (decoy database).
  - Protein from decoy database is false one.

Validation

Given a certain probability threshold t, F_t is the number of false positives,

- Fasle Discovery Rate (FDR):

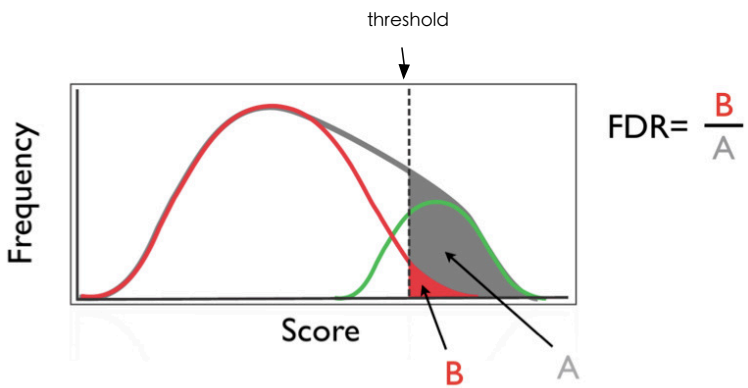
$$FDR_t = \frac{F_t}{F_t + T_t}.$$

- q-values:

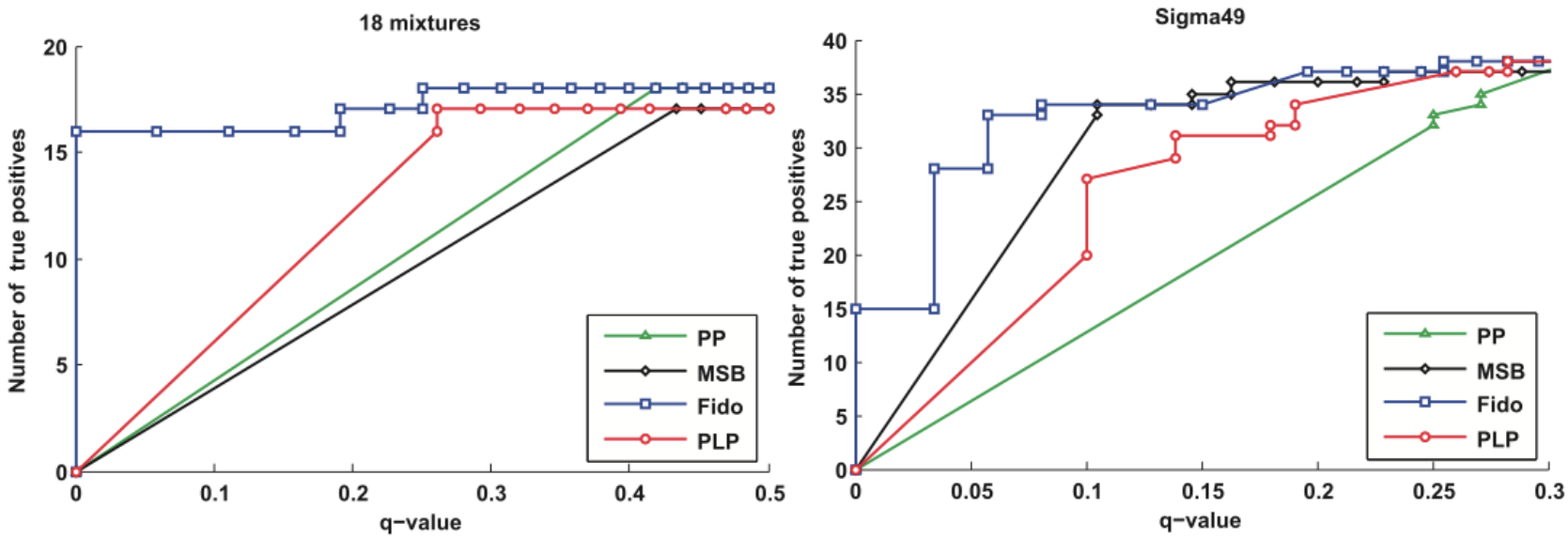
$$q_t = \min_{t' \leq t} FDR_{t'}.$$

- Posterior error probability (PEP):

$$PEP = Pr(+|p).$$



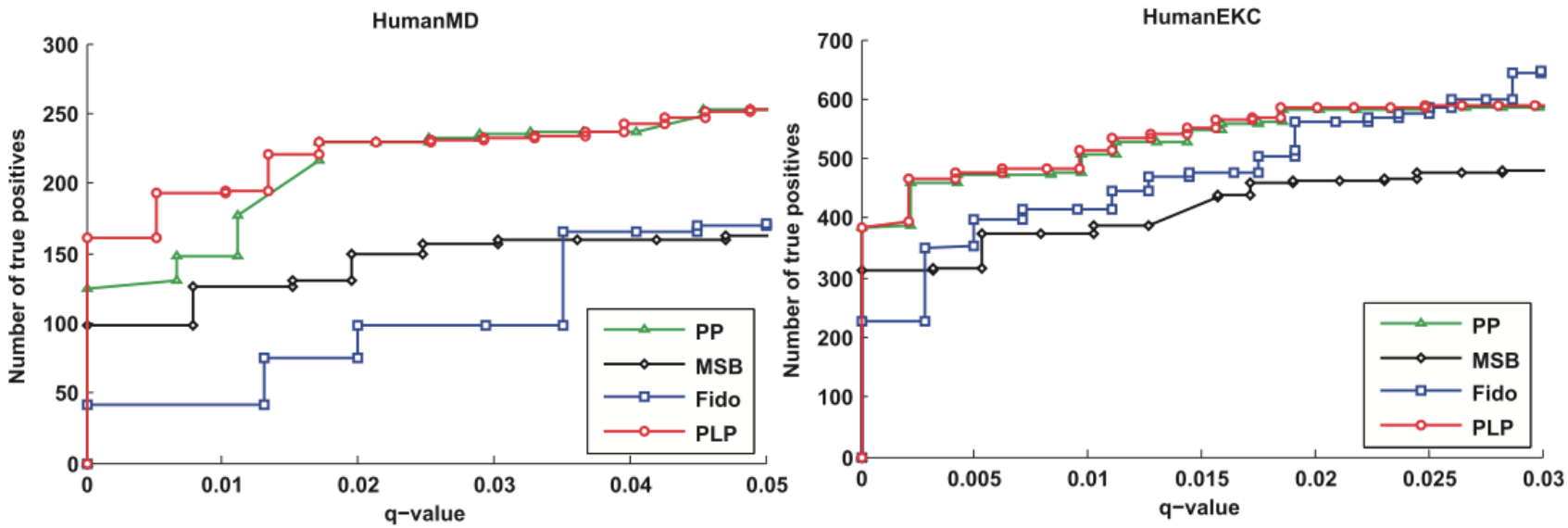
Comparison of q-values



MSB is MSBayespro[5].

[5] Yong Fuga Li et al. "A Bayesian approach to protein inference problem in shotgun proteomics". In: *Journal of Computational Biology* 16.8 (2009), pp. 1183–1193.

Comparison of q-values



Comparison: the number of degenerate peptides

	PP		ProteinLP		Fido	
	TP	FP	TP	FP	TP	FP
			18 mixtures			
Simple proteins	17	8	17	9	17	9
Degenerate Proteins	1	5	0	5	1	4
			Sigma49			
Simple proteins	27	1	30	1	30	5
Degenerate Proteins	5	10	5	7	5	3
			HumanMD			
Simple proteins	70	0	64	0	111	6
Degenerate Proteins	54	0	60	0	7	0

Table: Accuracy on proteins containing shared peptides with q-value threshold 0.3 for Sigma49 and 0.01 for HumanMD.

Part IV

Conclusions

Conclusions

- Joint probabilities provide the degeneracy information.
- Joint probabilities simplify the optimization problem.
- To do:
  - Integrate supplementary information, e.g. protein-protein interaction, by adding linear constraints.
  - Considering the parameter ϵ for different peptide probabilities and protein information.

Thanks for listening.

Questions?