Protein inference based on statistical modeling in bipartite graphs

Sarah Gerster

Seminar for Statistics, ETH Zurich

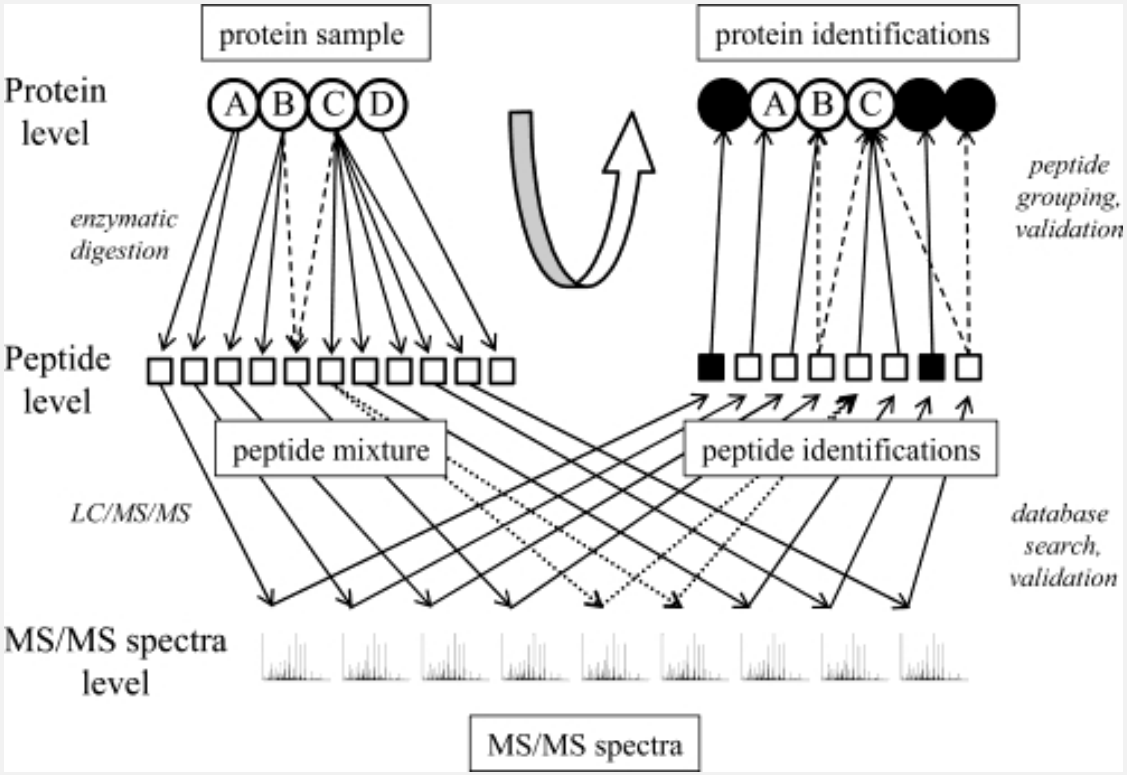
May 7, 2009

joint work with

Ermir Qeli, CMOP, University of Zurich
Christian Ahrens, CMOP, University of Zurich
Peter Bühlmann, SfS, ETH Zurich

protein inference

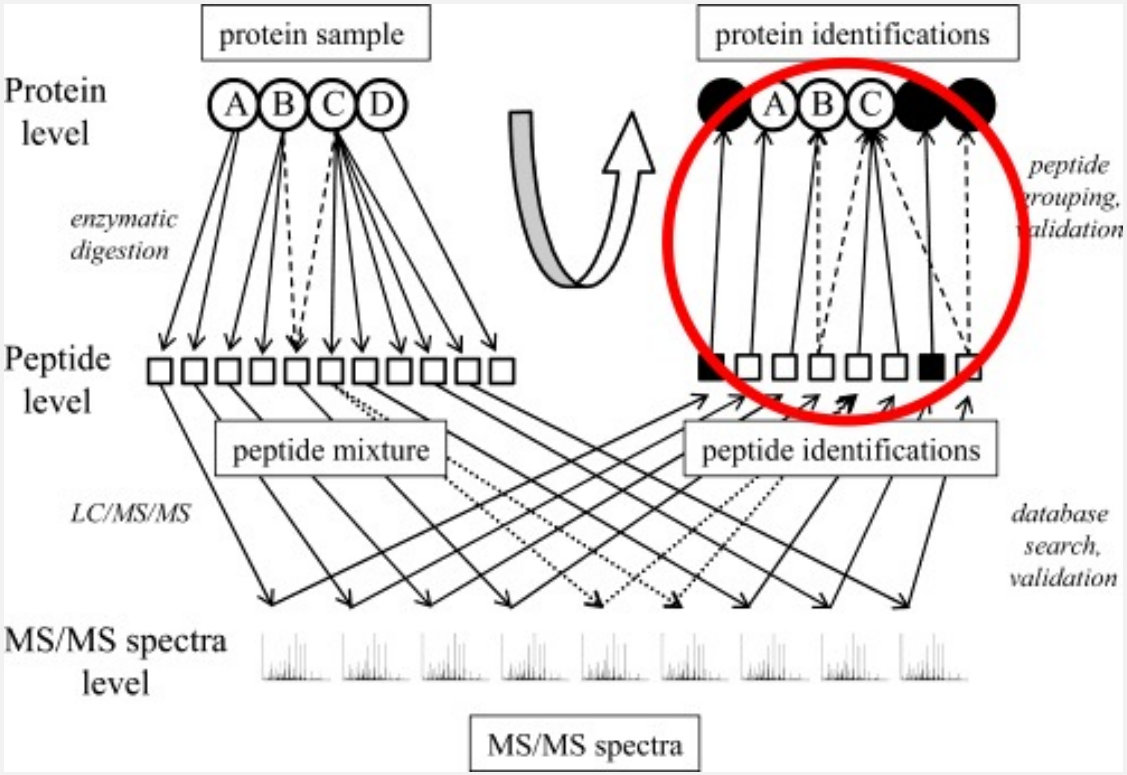
in a picture



(Nesvizhskii et al. 2003)

protein inference

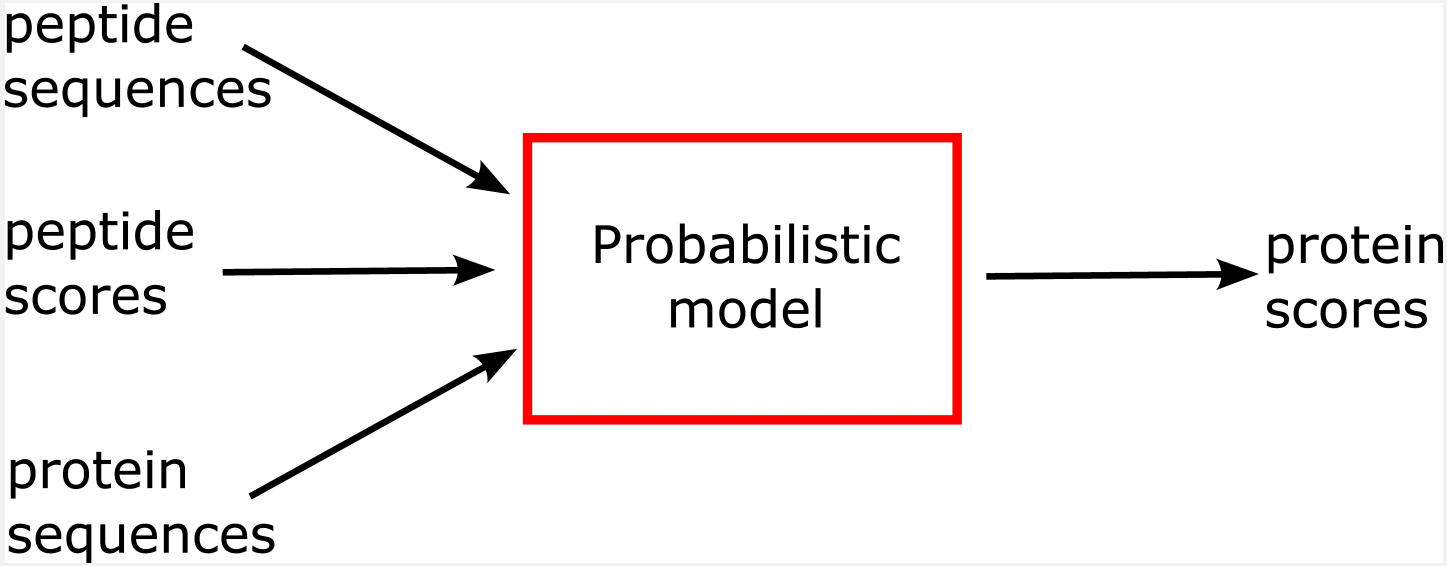
in a picture



(Nesvizhskii et al. 2003)

protein inference

schema



protein inference

Input:

- peptide identifications and scores from PeptideProphet (Keller et al. 2002)
- list of possible proteins in the sample:
  - proteins with at least one matching peptide
  - "minimal set" of proteins explaining all the peptides

Goal:

- score for each protein
- decide which proteins are in the sample

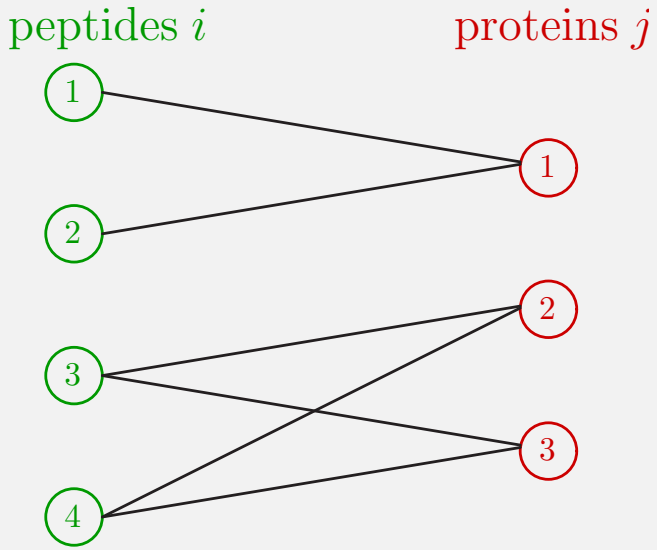
Applications:

- proteome annotation
- identification of proteins associated with a disease

bipartite graph

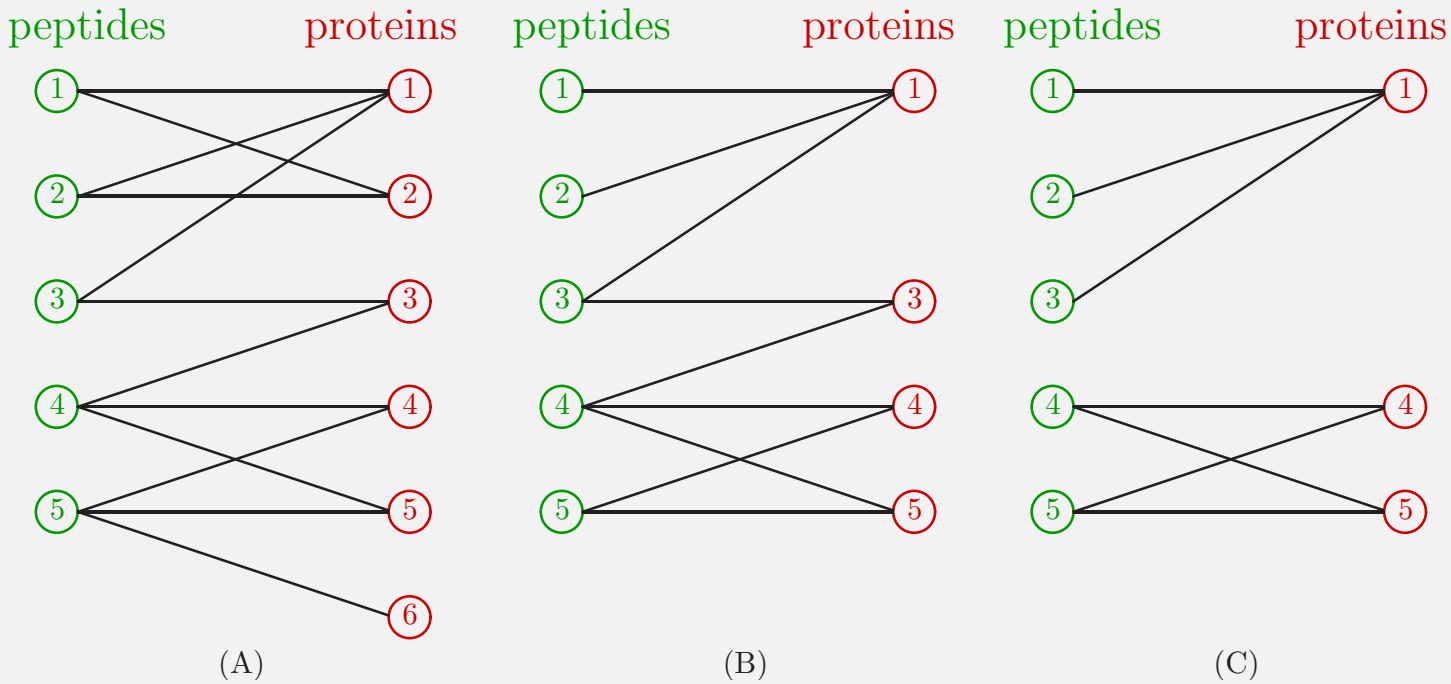
components

- 2 types of nodes:
  peptide and protein sequences
- edges between peptides and proteins
- edge:
  peptide sequence matches the protein sequence



bipartite graph

construction



peptides

assumptions and modeling

Assumptions:

- peptide scores are realisations of random variables

Implications:

- peptide scores are modeled by probability distributions
- uncertainty of peptide scores propagates to the protein scores

to be computed

$$\mathbb{P}[Z_j = 1 | \{p_i; i \in \mathcal{I}\}]$$

where

- Z_j indicates if protein j is present (1 stands for present, 0 for absent)
- p_i is the score of peptide i
- $\mathcal{I}$ is the set of all experimentally identified peptides

assumptions and modeling

Look at the problem the other way around by using Bayes' theorem:

- $\mathbb{P}[A|B] = \frac{\mathbb{P}[B|A]\mathbb{P}[A]}{\mathbb{P}[B]}$ = probability of protein presence given the peptide scores
- $\mathbb{P}[A]$ = protein prior
- $\mathbb{P}[B]$ = peptide probabilities
- $\mathbb{P}[B|A]$ = peptide probabilities given the presence or absence of proteins

connected components

assumptions and modeling

Assumptions:

- different connected components are independent
- peptides in the same connected component are independent given their neighboring proteins

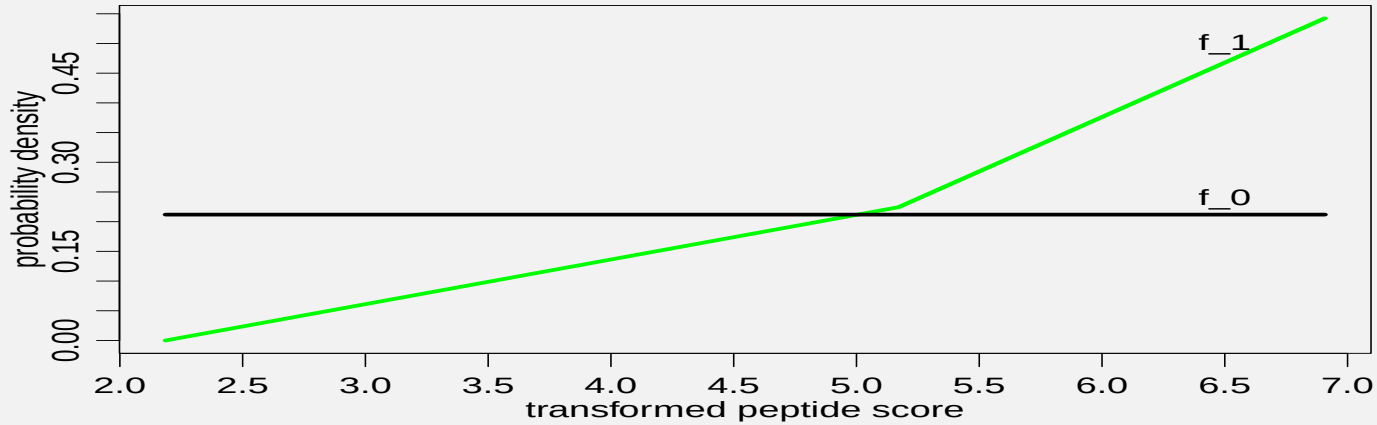
Implications:

- only have to look at one connected component at a time (not at the whole graph)
- the probability of a certain peptide score depends only on the peptide's neighboring proteins

mixture model

assumptions and modeling

$$\mathrm{p}\left(p_i | \{z_j; j \in Ne(i)\}\right) \sim \begin{cases} f_0(p_i) & \text{if } \sum_{j \in Ne(i)} z_j = 0 \\ f_1(p_i) & \text{if } \sum_{j \in Ne(i)} z_j > 0 \end{cases}$$



a special mixture model with 2 parameters

proteins

assumptions and modeling

Assumptions:

- prior presence of protein is independent of other proteins
- presence of protein is independent of the experimental conditions

In addition, we use:

- same prior probability for all proteins
  $\longrightarrow$ potential loss of biological knowledge

computations

parameter estimation

- estimate the parameters of the peptide probability distribution: MLE
- protein "priors" are estimated: MLE
- handle large connected components: random sampling

testing the model

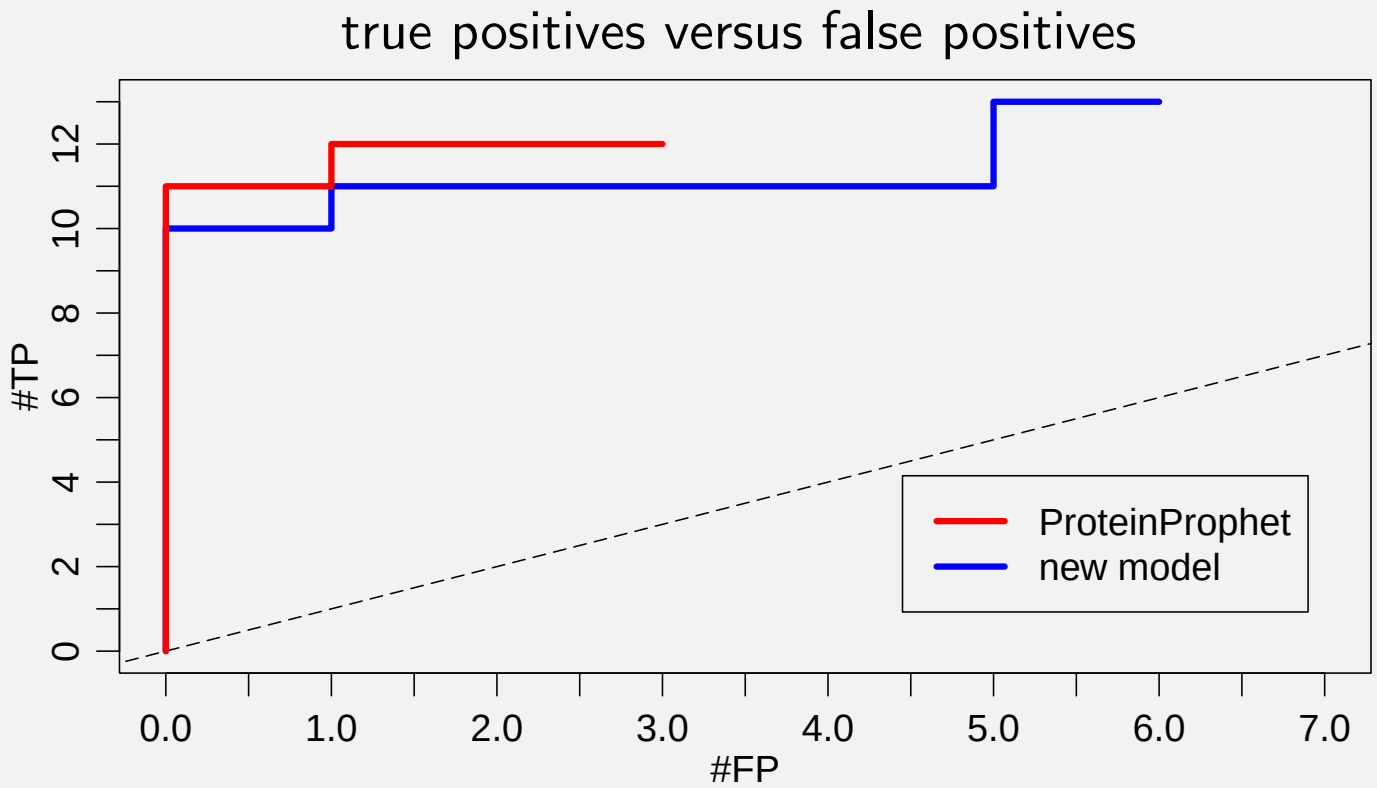
Tests on different datasets:

- two control datasets
  - mixture of 18 purified proteins (Keller et al. 2002)
  - Sigma49 (Tabb et al. 2007)
- real data
  - *Drosophila melanogaster* (Brunner et al. 2007)

Evaluation: comparison with ProteinProphet (Nesvizhskii et al. 2003)

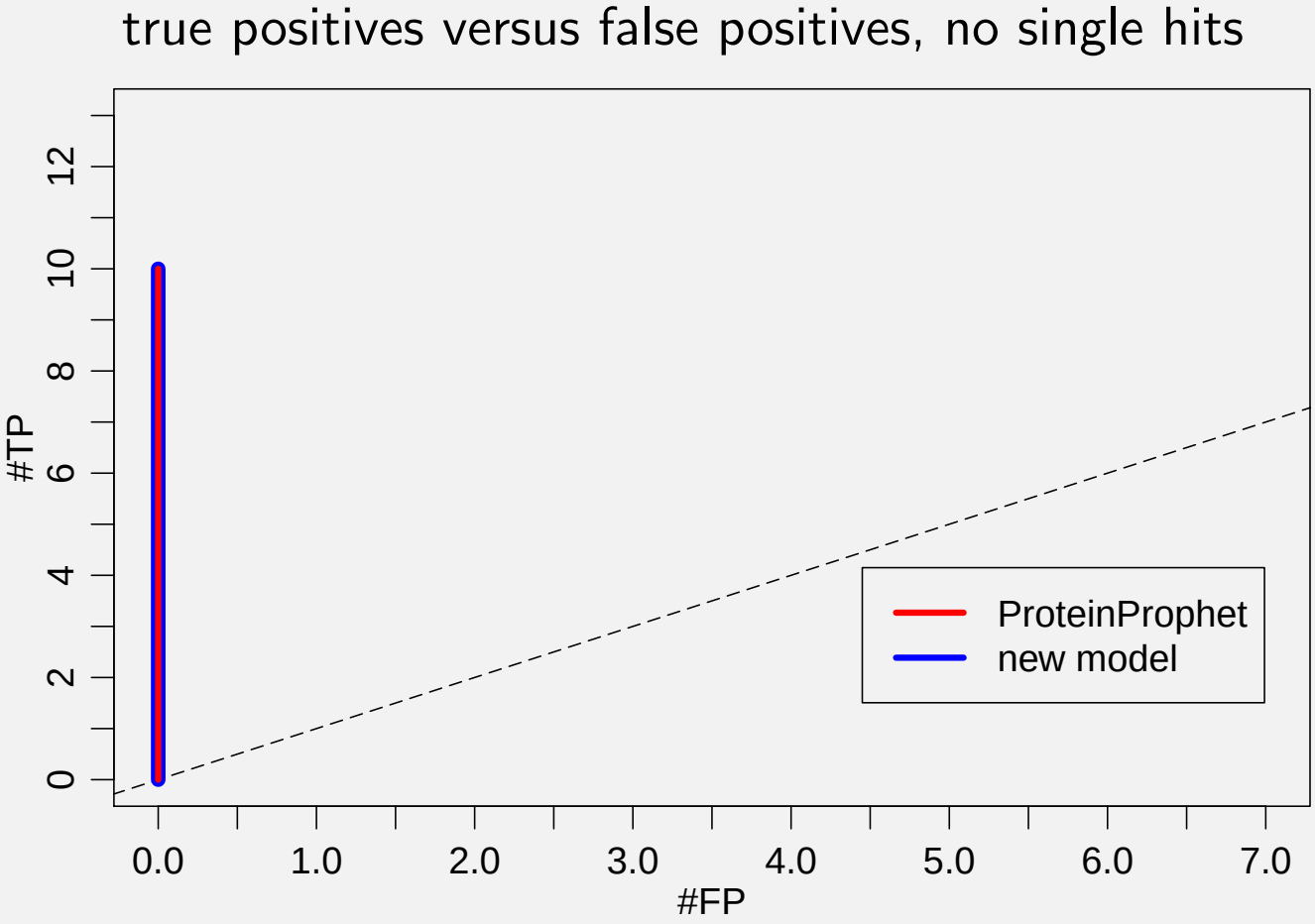
mixture of 18 purified proteins

control dataset



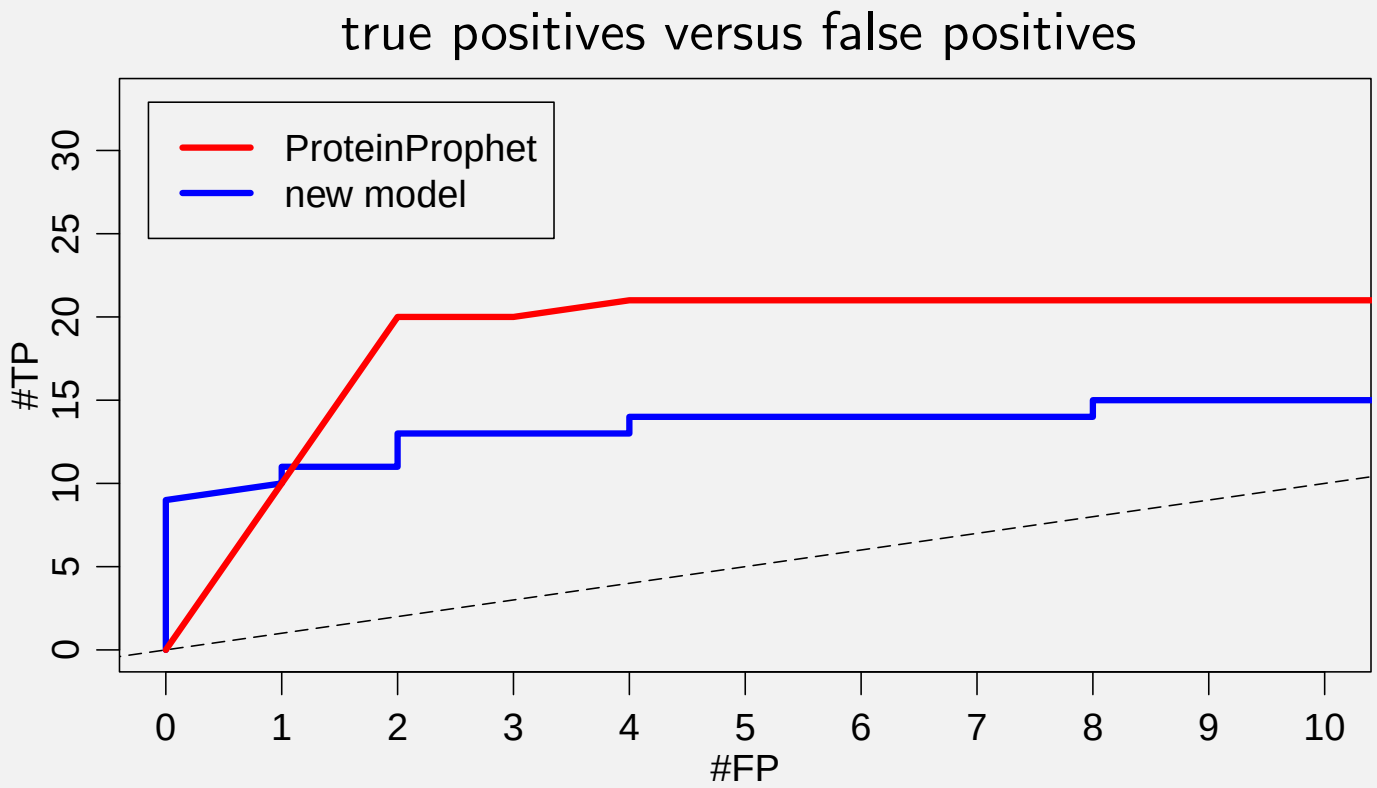
mixture of 18 purified proteins

control dataset



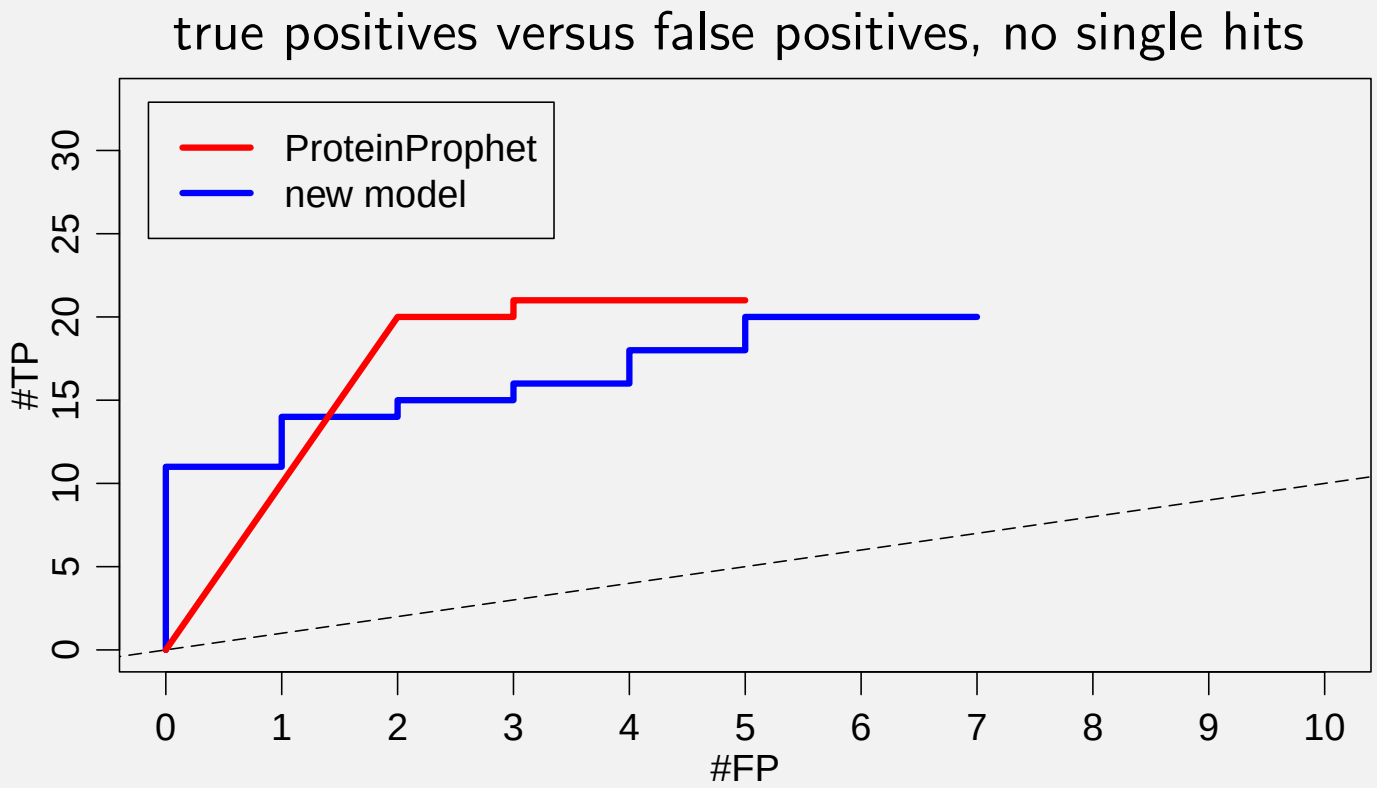
sigma49

control dataset



sigma49

control dataset



D. melanogaster

real data

Including single hits:

n	25	50	76	100	168	205	219
intersection	25	50	76	100	113	138	152

Not including single hits:

n	25	50	76	100	150	200	219
intersection	25	50	76	100	123	169	179

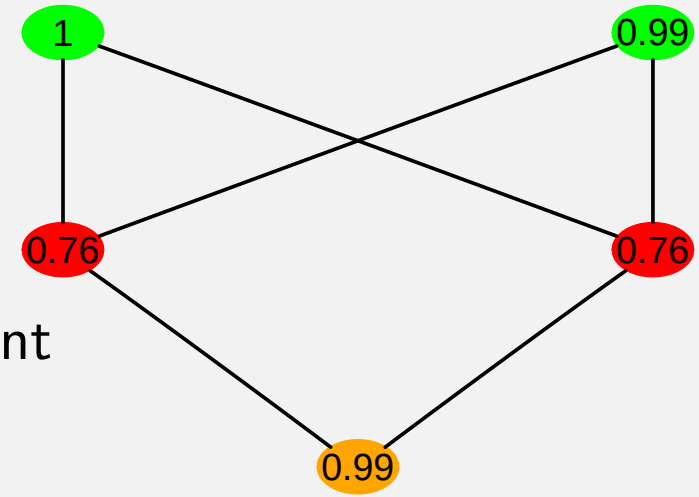
D. melanogaster

real data

application to gene models

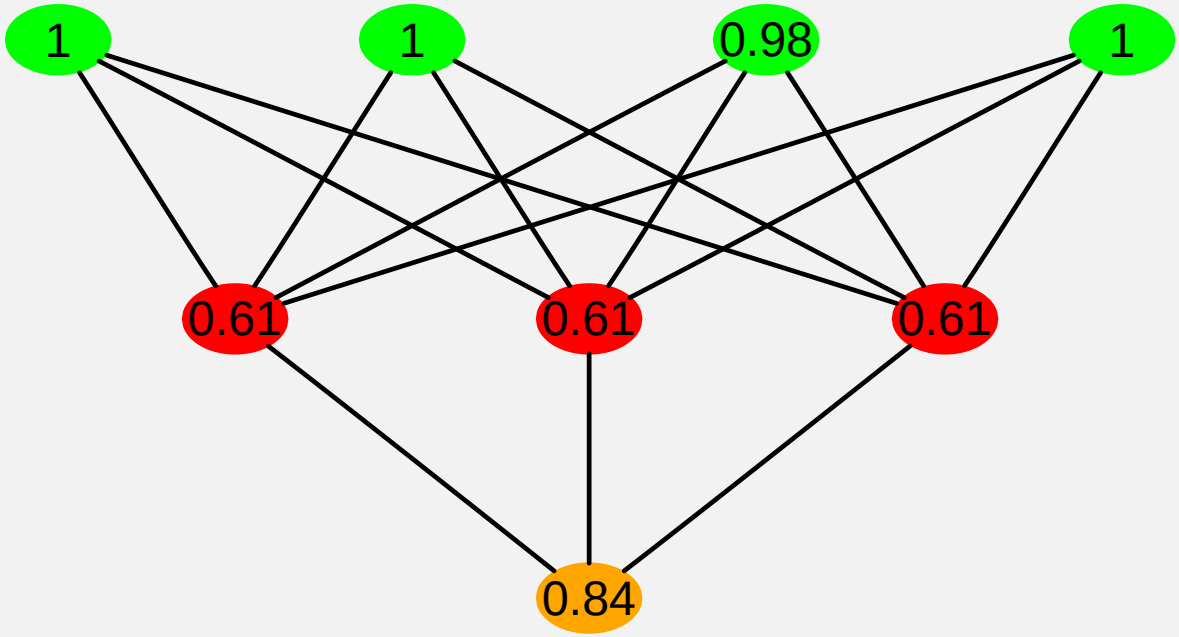
A new level is added to the graph: "tripartite" graph

- shared peptides
- proteins may not be clearly identifiable
- several proteins from the same gene model
- try to make a statement about the gene model



modeling for gene models

$$\mathbb{P}[\text{gene model occurs}] = 1 - \mathbb{P}[\text{none of its proteins occur}]$$



The three proteins here are CG12013-PA, CG12013-PC and CG12013-PD.

summary

We present a new model with

- peptide probabilities modeled as random quantities
- transparent uncertainty propagation from the peptide level to the protein level
- an extension to compute probabilities of a gene model being present or not in the sample

Our results look promising when compared to ProteinProphet.

references

- Nesvizhskii A I, Keller A, Kolker E, Aebersold R (2003) A statistical model for identifying proteins by tandem mass spectrometry. *Anal. Chem.* 75:4646-4658.
- Keller A, Nesvizhskii A I, Kolker E, Aebersold R (2002) Empirical statistical model to estimate the accuracy of peptide identifications made by ms/ms and database search. *Anal. Chem.* 74:5383-5392.
- Keller A, Purvine S, Nesvizhskii A I, Stolyar S, Goodlett D R, Kolker E (2002) Experimental protein mixture for validating tandem mass spectral analysis. *OMICS: A Journal of Integrative Biology* 6:207-212.
- Tabb D L, Fernando C G, Chambers M C (2007) Myrimatch: Highly accurate tandem mass spectral peptide identification by multivariate hypergeometric analysis. *Journal of Proteome Research* 6:654-661.
- Brunner E, Ahrens C H, Mohanty S, Baetschmann H, Loevenich L, Potthast F, Deutsch E W, Panse C, de Lichtenberg U, Rinner O, Lee H, Pedrioli P G A, Malmstrom J, Koehler K, Schrimpf S, Krijgsveld J, Kregenow F, Heck A J R, Hafen E, Schlapbach R, Aebersold R (2007) A high-quality catalog of the *Drosophila melanogaster* proteome. *Nature Biotechnology* 25:576-583.