Peptide Assignment Validation

Telling what's wrong without actually knowing what's right

Patrick Pletscher

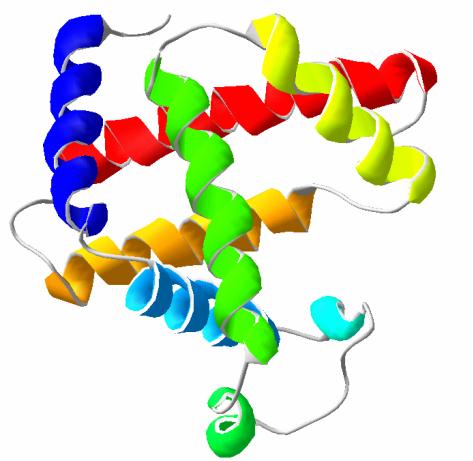
ETH Zurich, Switzerland

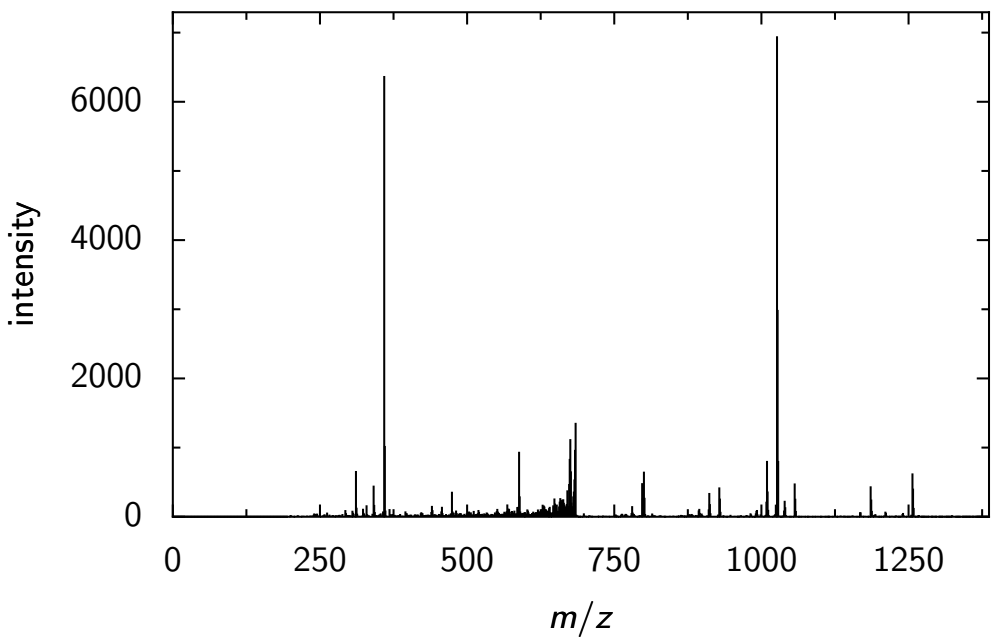
4th April 2006

Overview

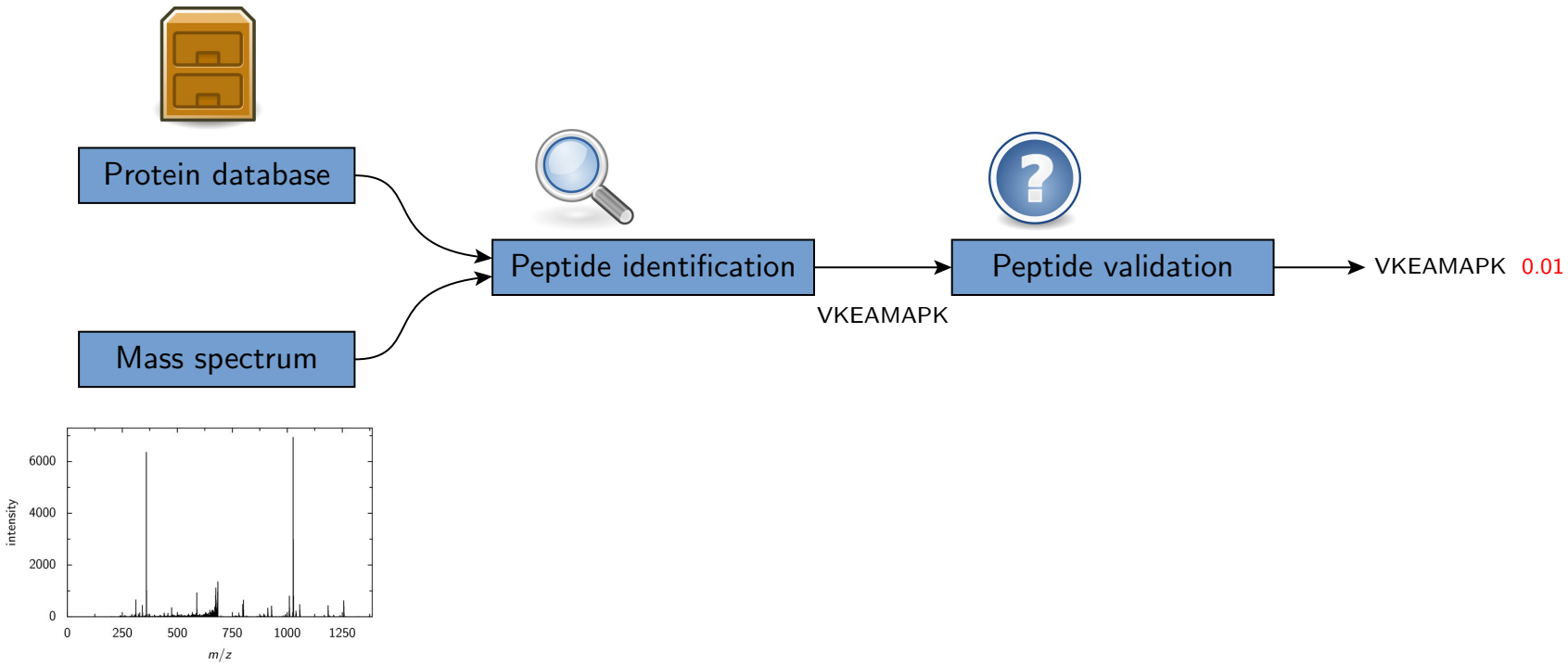
The Big Picture

- Given a cell: Which proteins are in it?
- Chemical processes, to reduce the complexity.
- In the end we get a mass spectrum of a peptide.
- Question: which *peptide* is it?





Question: How to validate DB searches?



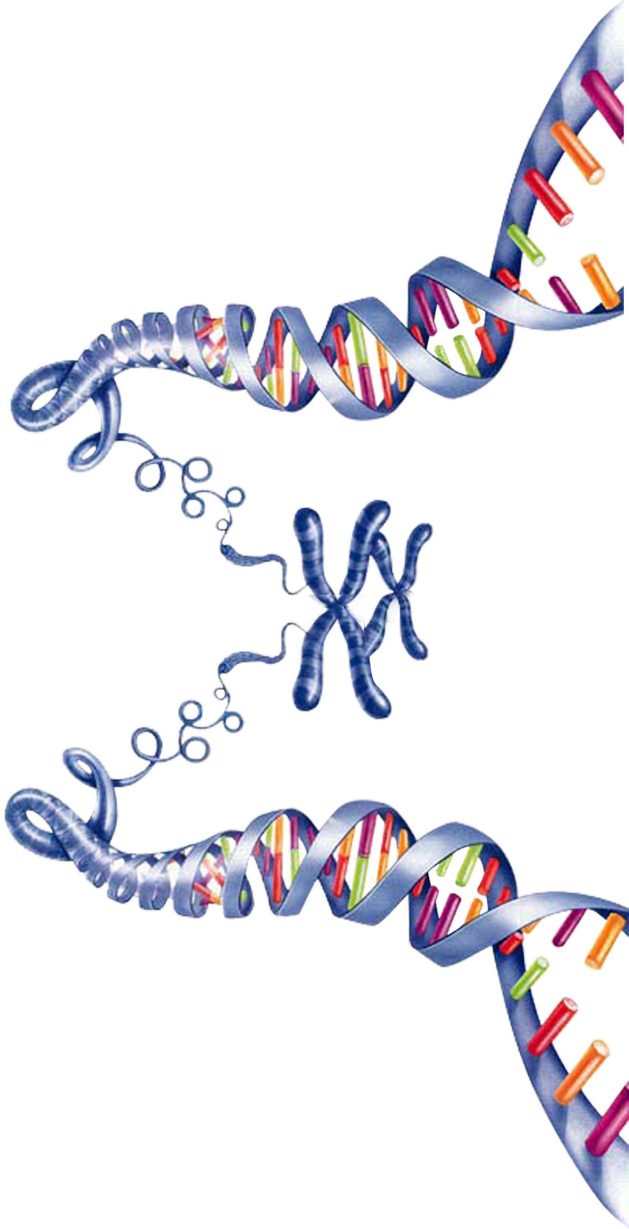
Goal of semesterthesis

Implement/test/extend database validation algorithms.

Protein Databases

Do you remember this thing called DNA?

- Protein information stored in the DNA.
- Genome of many organisms has been sequenced.
- Lots of data, we can use ...



Theoretical spectra

- Only sequences stored in DB.
- Split protein sequences to get peptides.
- Generate a theoretical spectrum for each DB entry.

A simple Algorithm

1. Split proteins in DB in all possible peptides.
2. Generate spectrum for each peptide and compare with experimental spectrum.
3. Return most similar sequence.
4. Validate the peptide assignment.

Data Normalization (1/2)

Flatten Experimental Spectrum

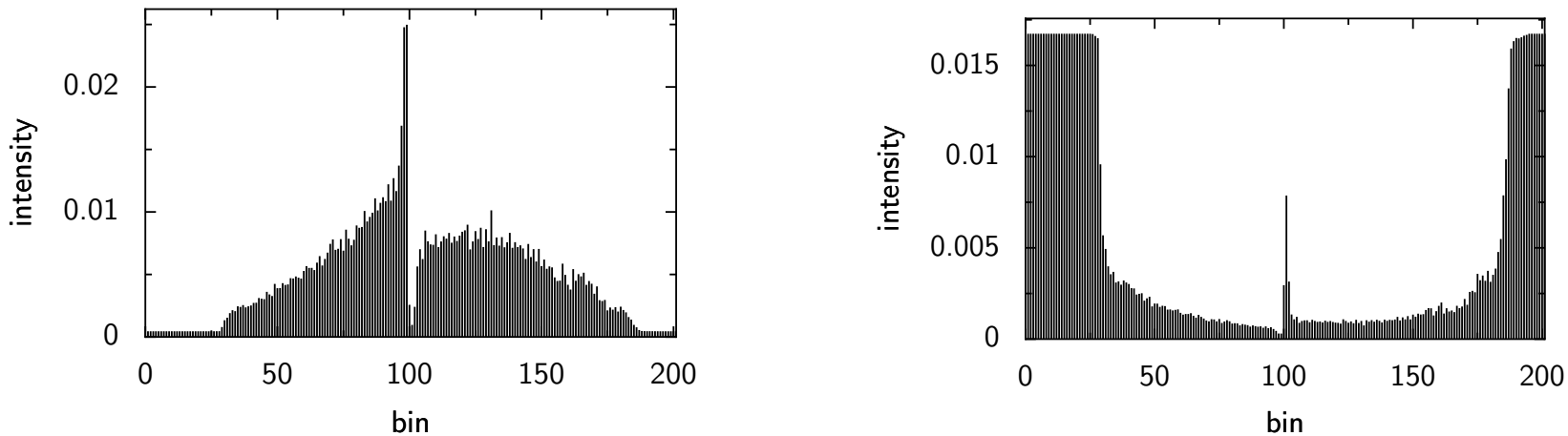


Figure: Learned intensity distribution (left) and inverse of it (right).

Data Normalization (2/2)

Remove Noise and Binarize

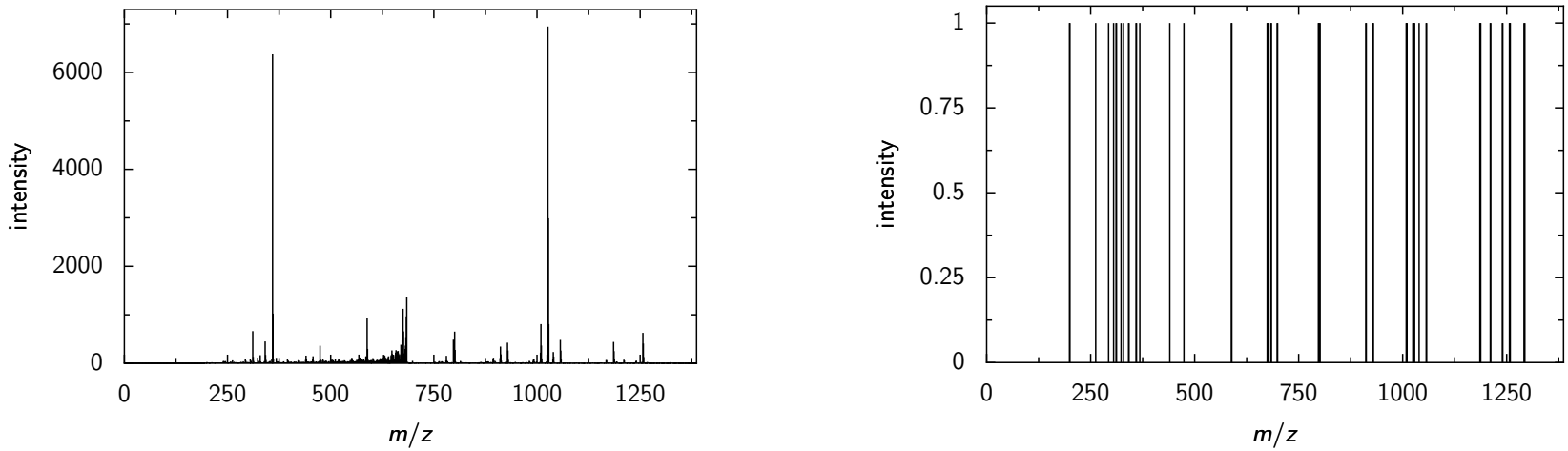
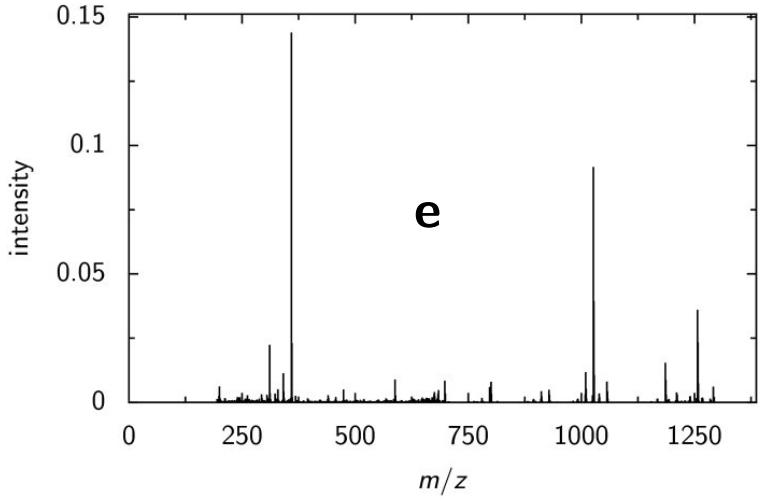
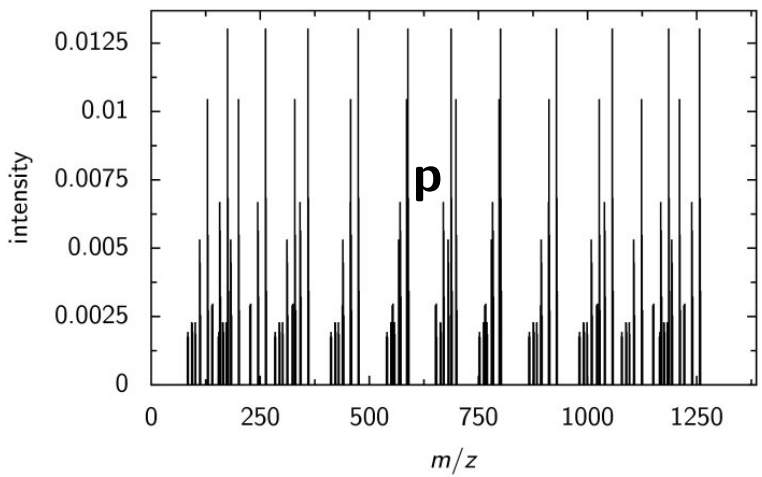


Figure: Experimental spectrum (left) and thinned out, binarized spectrum (right).

Cross-Correlation



experimental spectrum



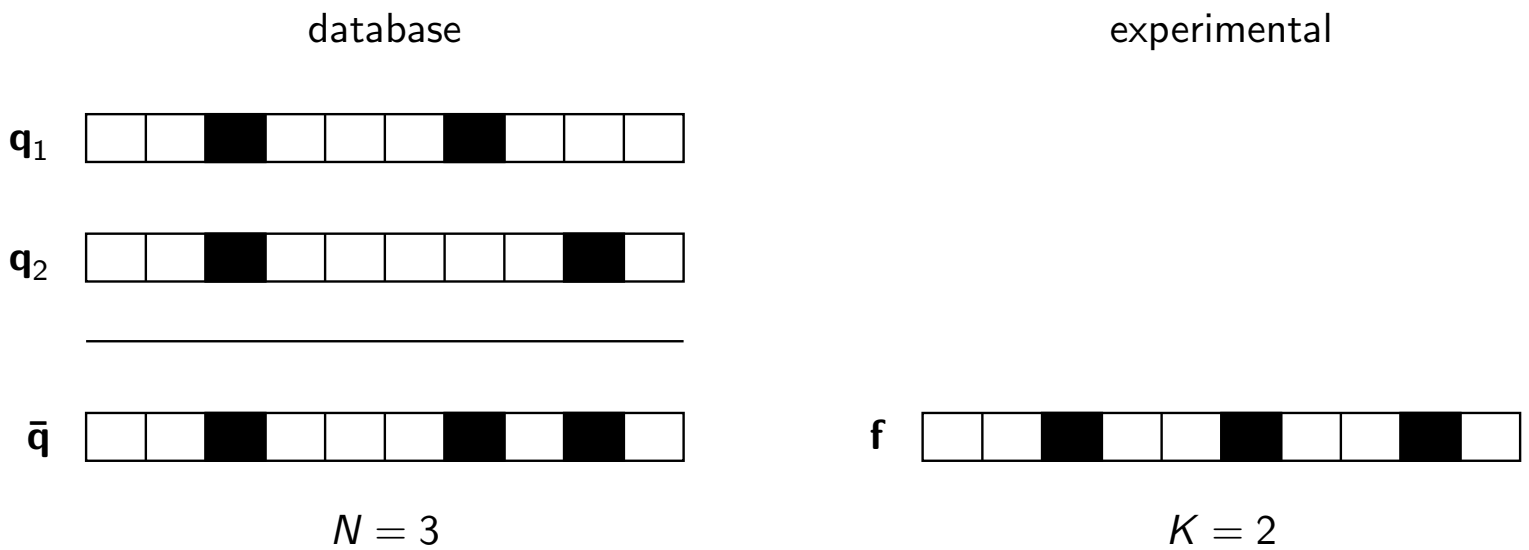
theoretical spectrum

$$Xcorr = \mathbf{e}^t\mathbf{p}$$

Hypergeometric Probability Model

Idea: What's the probability that the peptide sequence is a *random match*?

$$P_{N,K,n}(X = k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}$$



Validation – Different approaches

Goal

Decide whether assignment is correct.

Wide range of machine learning techniques can be applied:

- Statistical Tests, e.g. p-values.
- Discriminant Analysis between bad and good assignments (supervised).
- Gaussian Mixture Model (unsupervised).

Hypergeometric *p*-value

Statistical test if assignment is random.

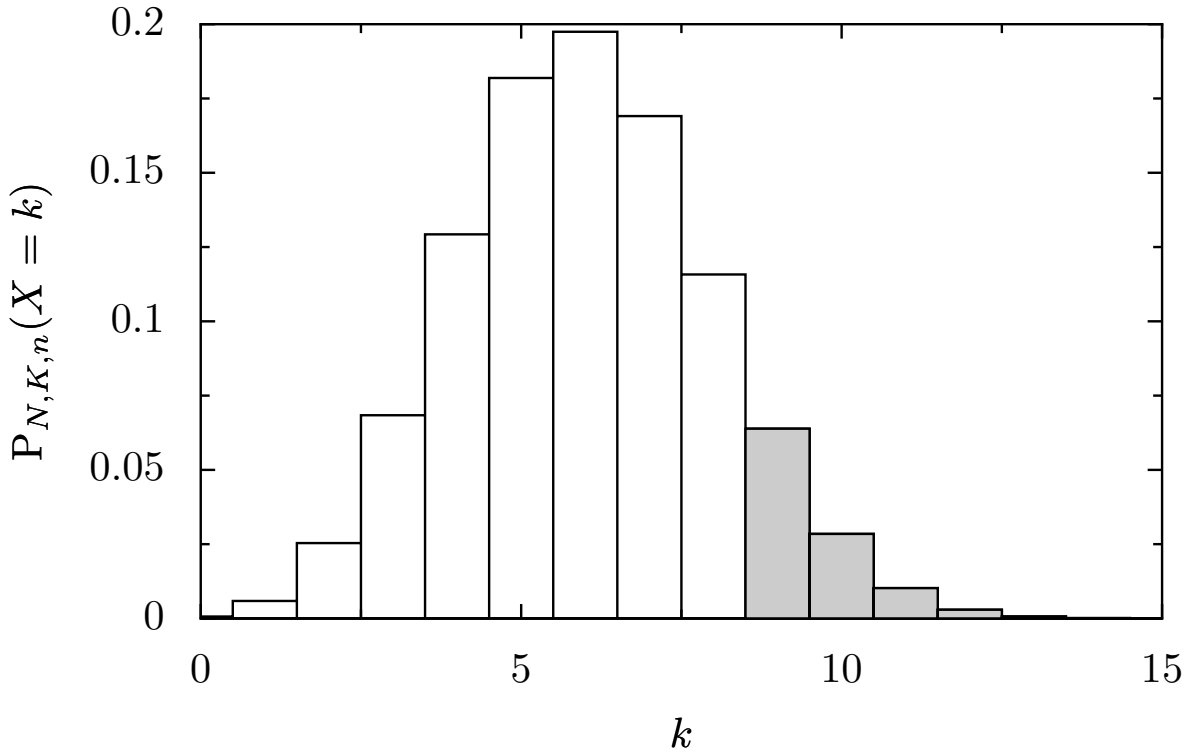


Figure: Hypergeometric probability distribution for $N = 1500, K = 500$ and $n = 18$ and p-value for $k = 9$ (shaded area).

Advantage: no training and thus no labels needed!

LDA validation

Discriminate between good and bad assignments.

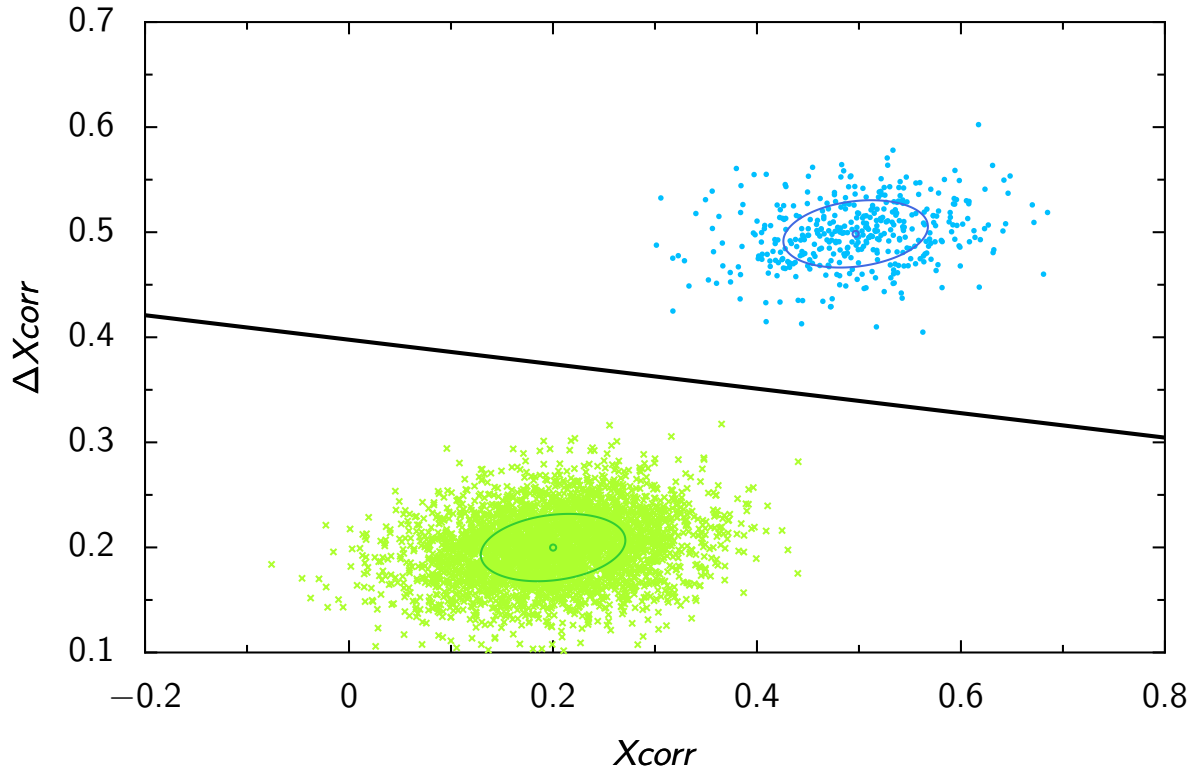
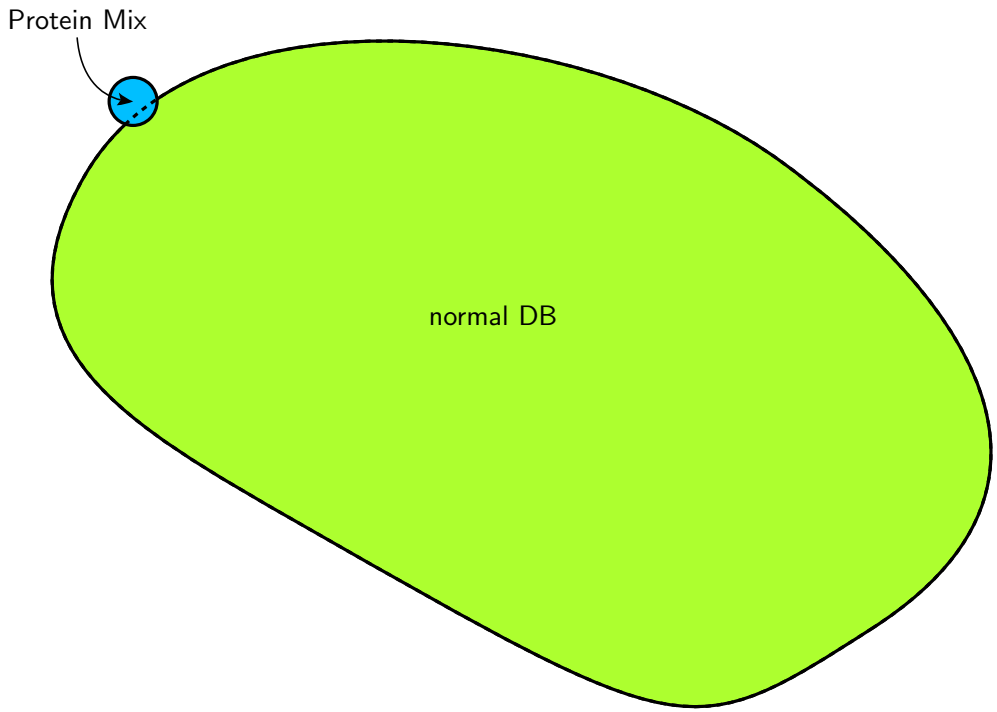


Figure: LDA example for artificial, normally distributed data.

Note: Needs labels!

Getting to know the labels – Protein Mix



Labels almost certainly correct, however expensive to get such data.

Getting to know the labels – Inverse DB

Idea: Two DBs: Containing both, normal and reverted proteins. If assignment from normal: correct, otherwise: wrong.

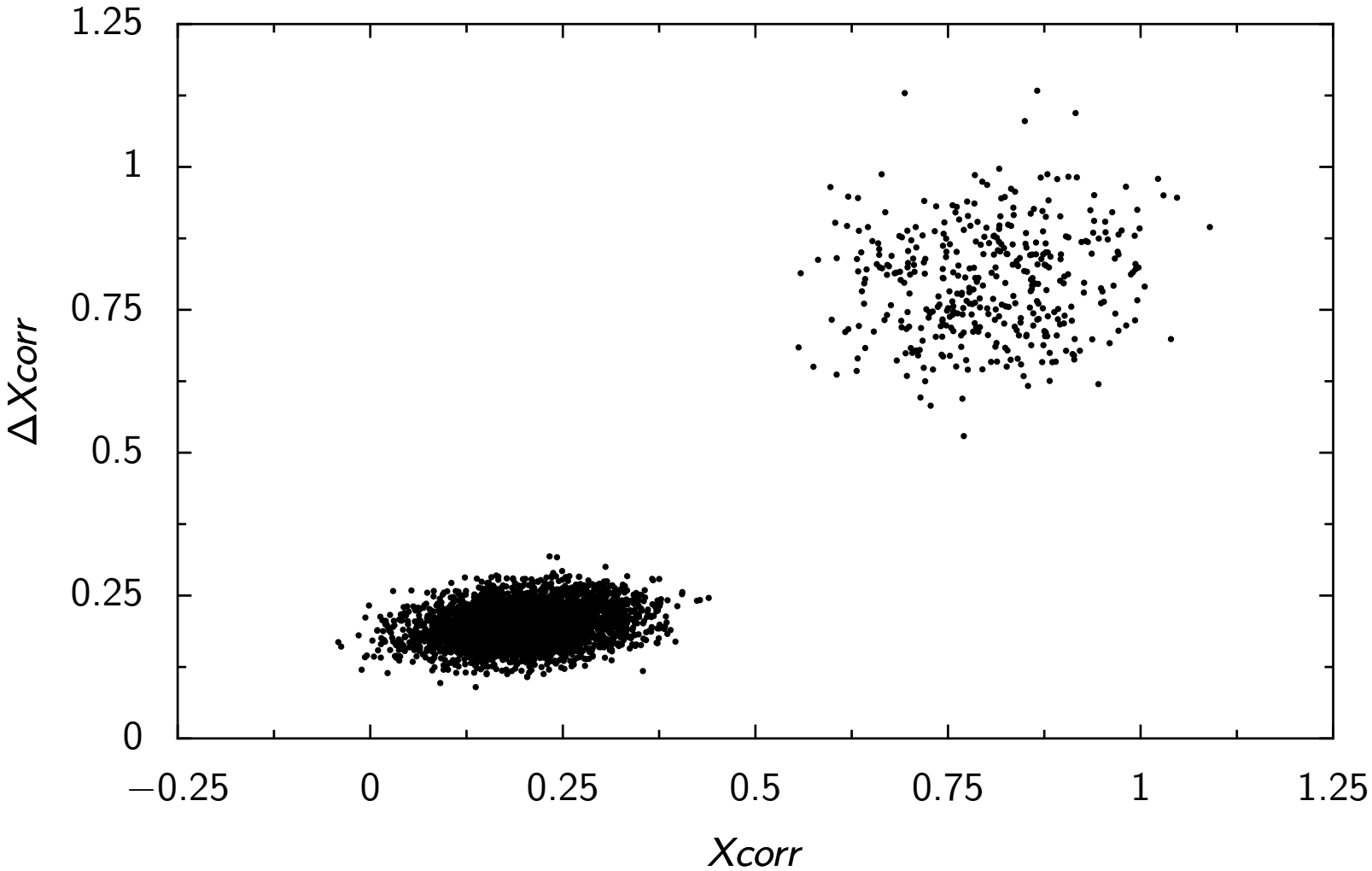
MEDQVGF...SWIILVG

$\forall$ proteins $\in$ DB

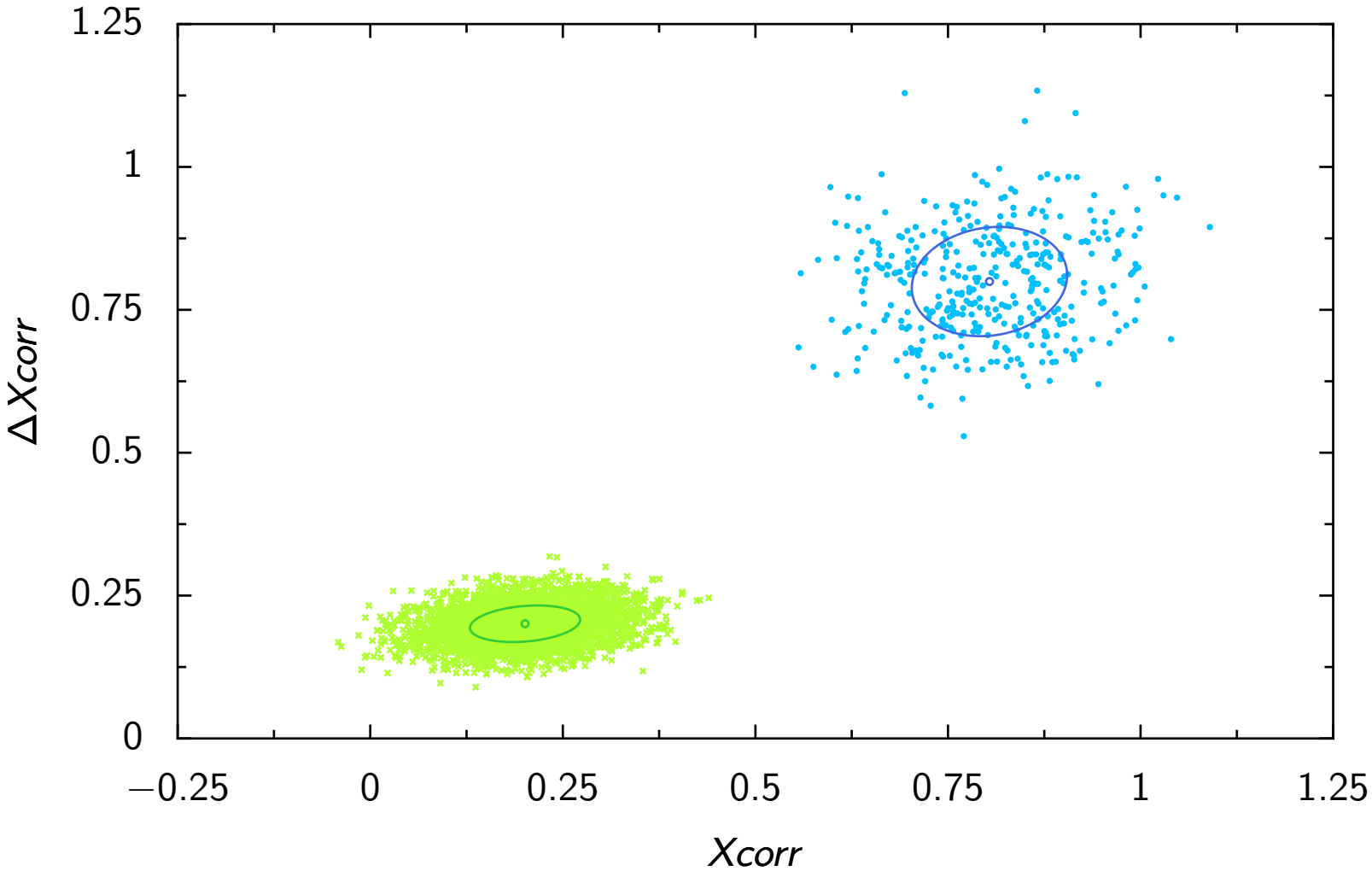
GVLIIWS...FGVQDEM

Lot of randomness: assignments from normal DB not guaranteed to be correct.

Gaussian Mixture Model (EM Algorithm)



Gaussian Mixture Model (EM Algorithm)



ROC and Precision-Recall

ROC: usually used in classification

$$sensitivity = \frac{tp}{tp + fn}$$

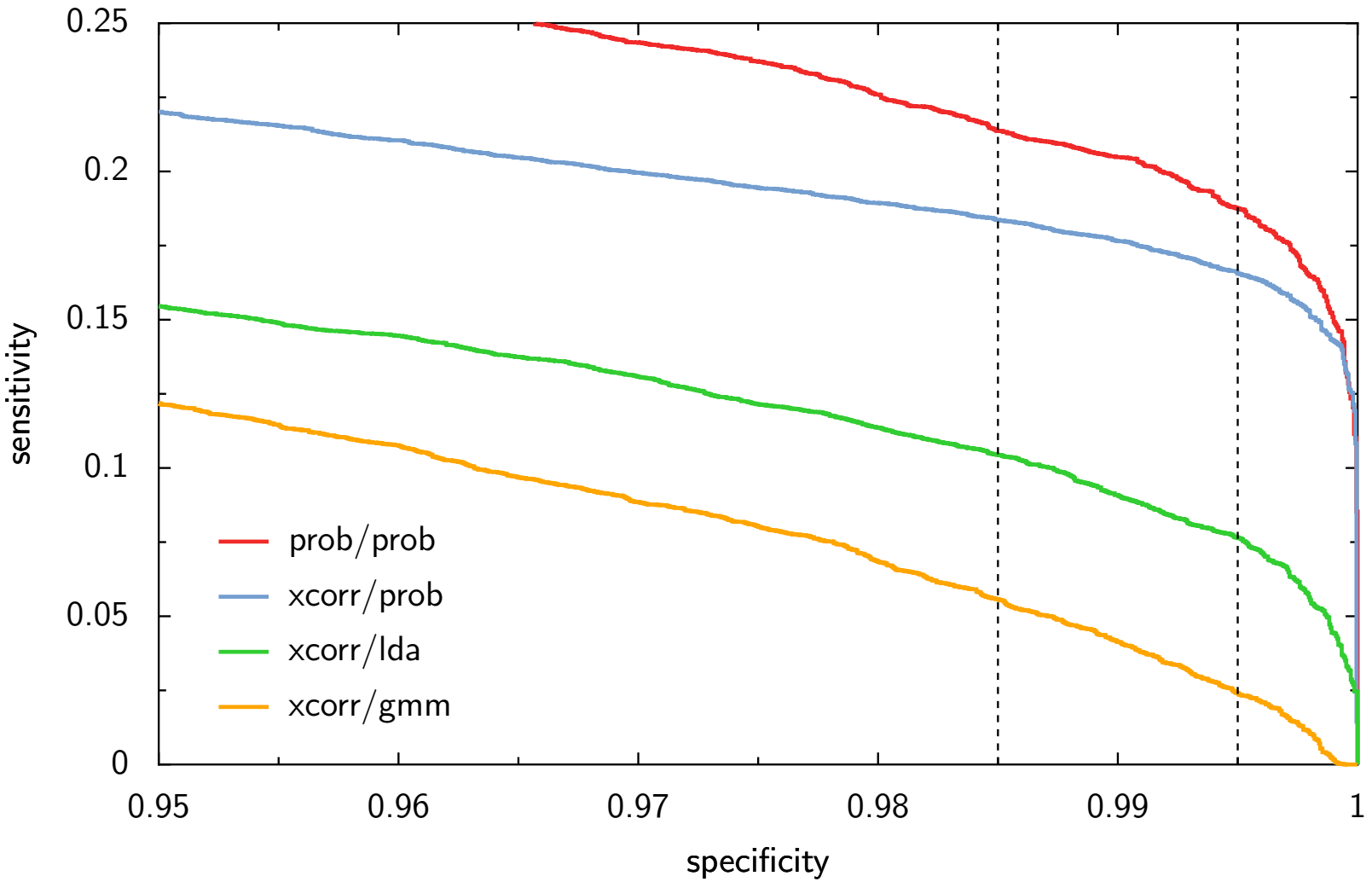
$$specificity = \frac{tn}{fp + tn}$$

Precision-Recall: usually used in information retrieval

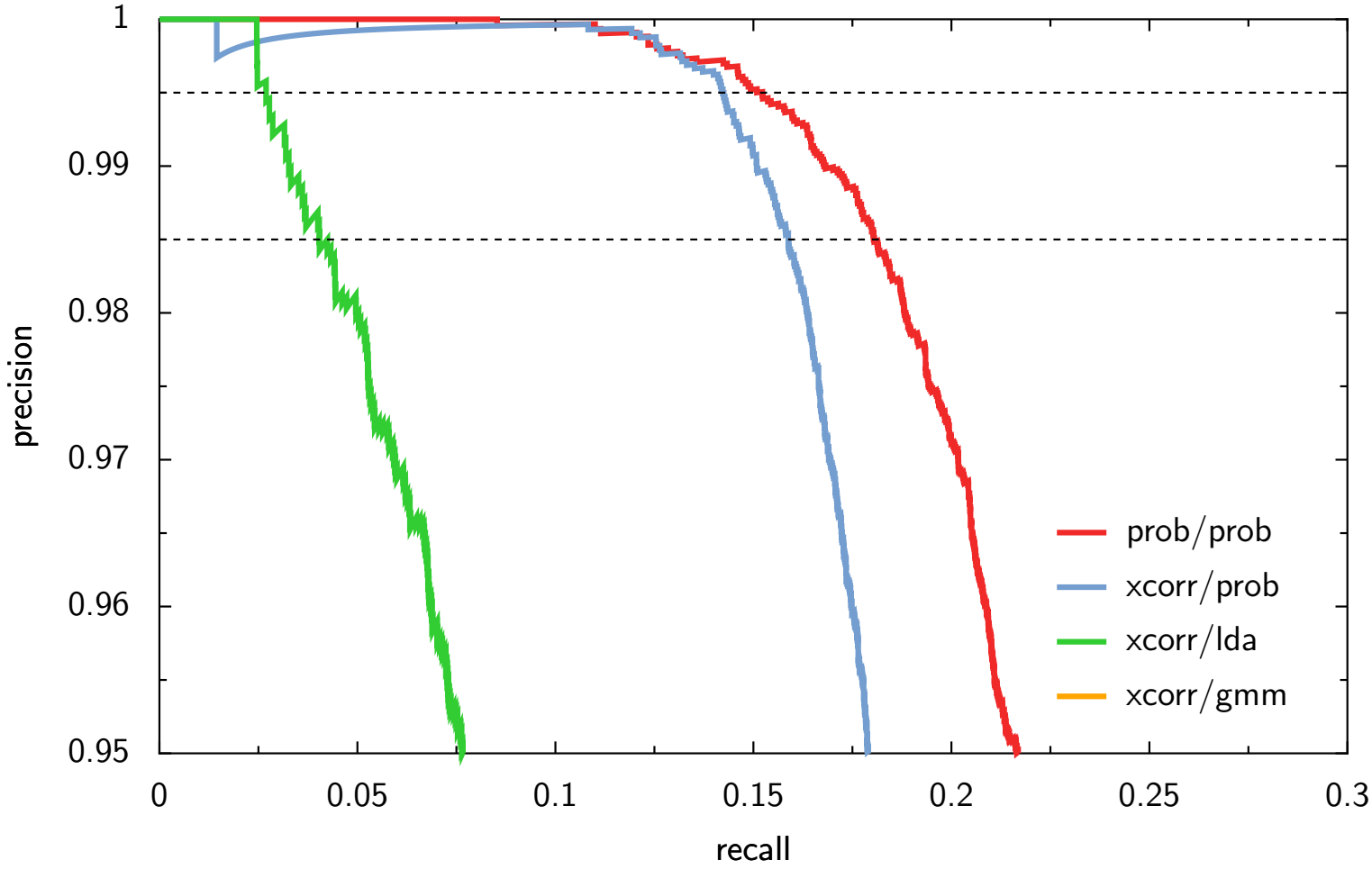
$$precision = \frac{tp}{tp + fp}$$

$$recall = \frac{tp}{tp + fn}$$

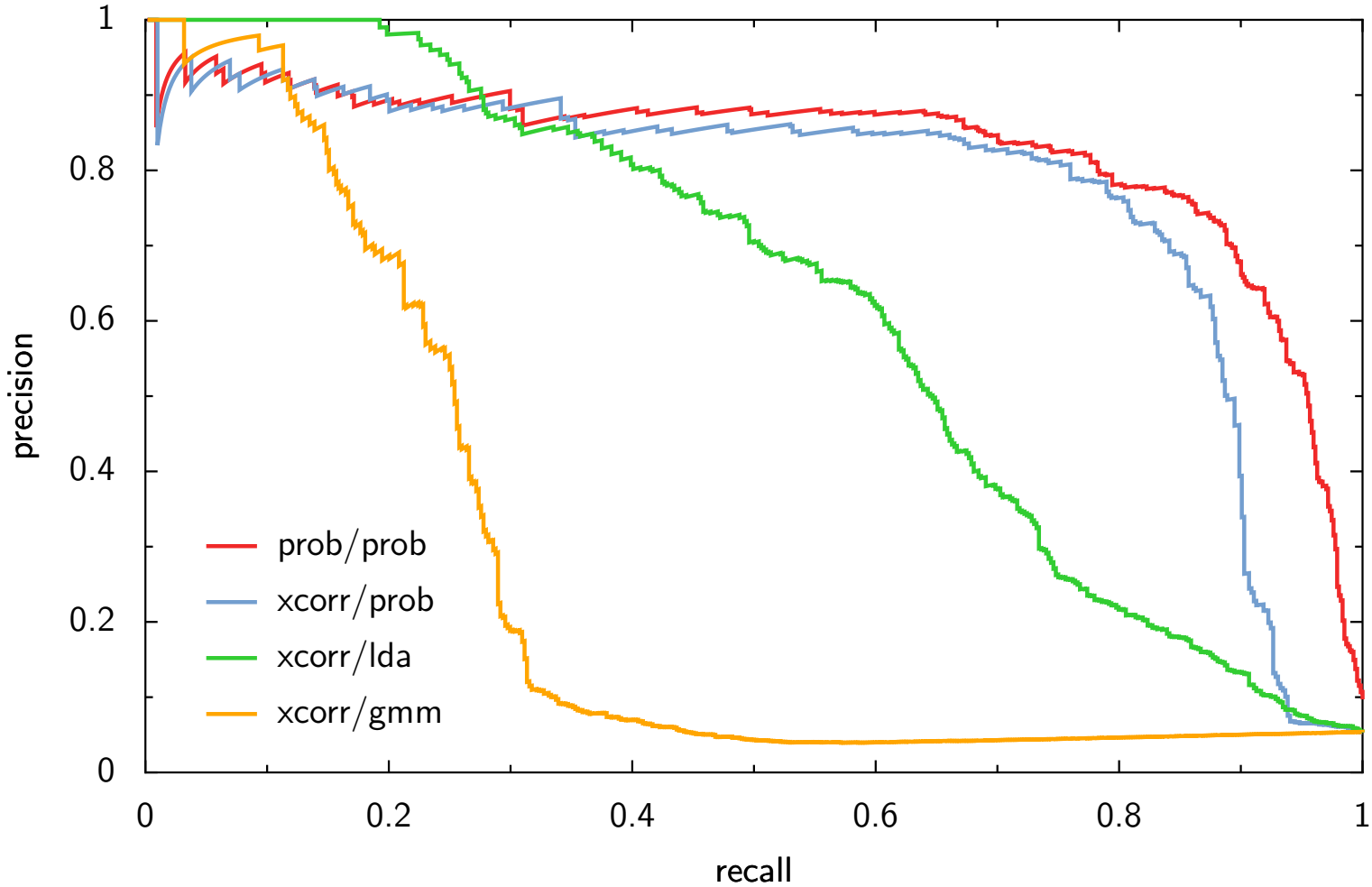
Evaluation with Inverse Database – ROC



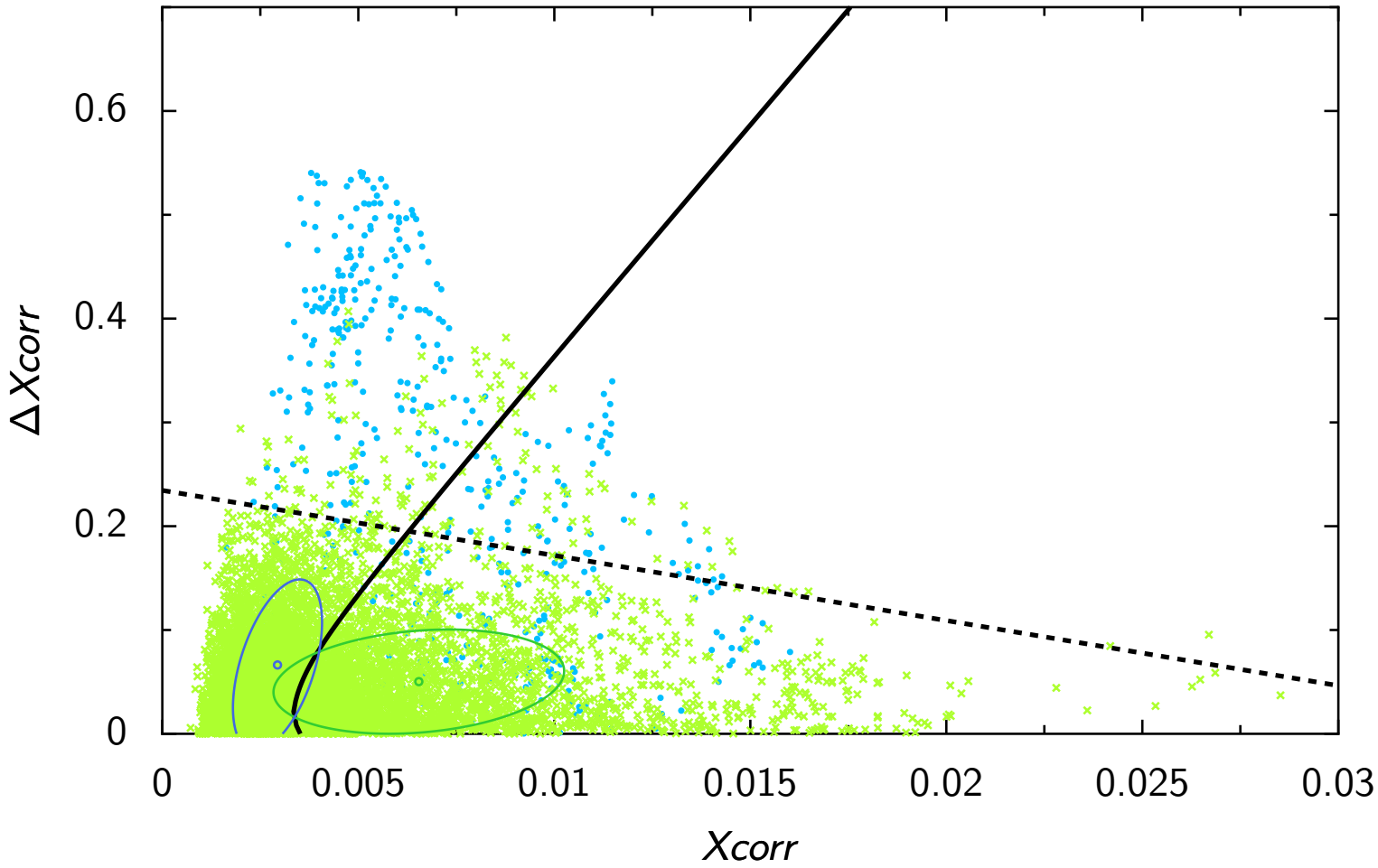
Evaluation with Inverse Database – Precision-Recall



Evaluation with Protein Mix – Precision-Recall



What's wrong with your GMM?



Conclusions

- *p*-value captures *some* information about correctness of assignment.
- However: further (exhaustive) tests needed to quantify its strengths and weaknesses.
- Test other classification algorithms, e.g. decision trees.
- Although LDA showed to be competitive, unclear why one should use classification approach.
- Random Probes of bad class easy to generate.
- Precision-Recall as measurement of choice.