Mass Spectra Alignments and their Significance

Sebastian Böcker[1], Hans-Michael Kaltenbach[2]

[1] Technische Fakultät, Universität Bielefeld

[2] NRW Int'l Graduate School in Bioinformatics and Genome Research, Universität Bielefeld

Overview

- Mass Spectrometry in Proteomics
- Protein Identification via MS
- Alignment of Spectra
- Score Significance
- Conclusion

Overview

- Mass Spectrometry in Proteomics
- Protein Identification via MS
- Alignment of Spectra
- Score Significance
- Conclusion

Overview

- Mass Spectrometry in Proteomics
- Protein Identification via MS
- Alignment of Spectra
- Score Significance
- Conclusion

Overview

- Mass Spectrometry in Proteomics
- Protein Identification via MS
- Alignment of Spectra
- Score Significance
- Conclusion

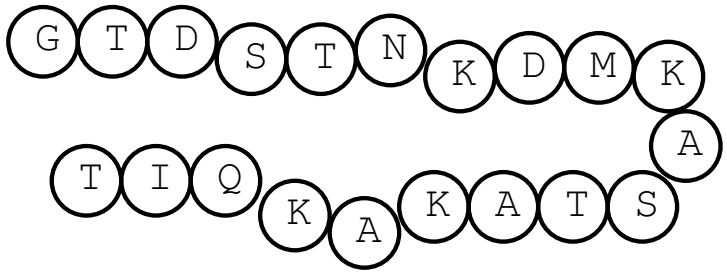
Overview

- Mass Spectrometry in Proteomics
- Protein Identification via MS
- Alignment of Spectra
- Score Significance
- Conclusion

Proteins

Biology

Proteins are directed polymers of 20 different amino acids.



Mathematics

Proteins are strings over an alphabet Σ.

Mass Spectrometry

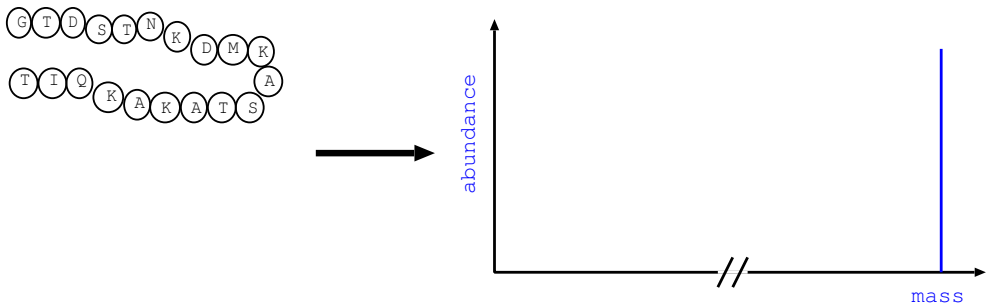
Mass Spectrometry in Bioscience

Mass spectrometry measures the masses and quantity of molecules in a probe.
It is widely used in biosciences to identify proteins and other biomolecules.

Fragmentation of peptides

Problem

Solely measuring the mass of a protein is not sufficient for identification.



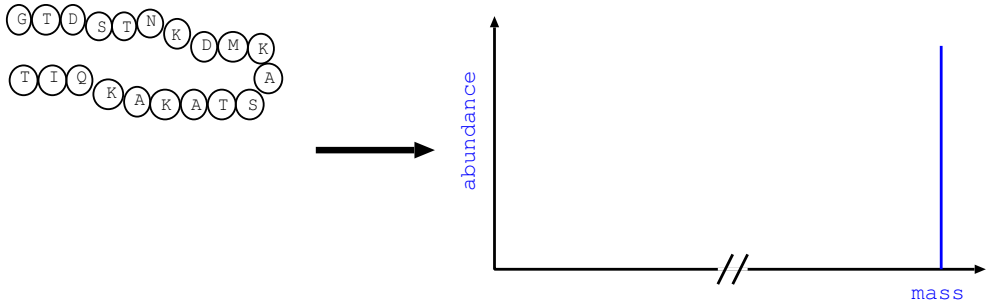
Idea

Break up the protein into smaller pieces in a deterministic way.
The spectrum of these pieces is called a *fingerprint* of the protein.

Fragmentation of peptides

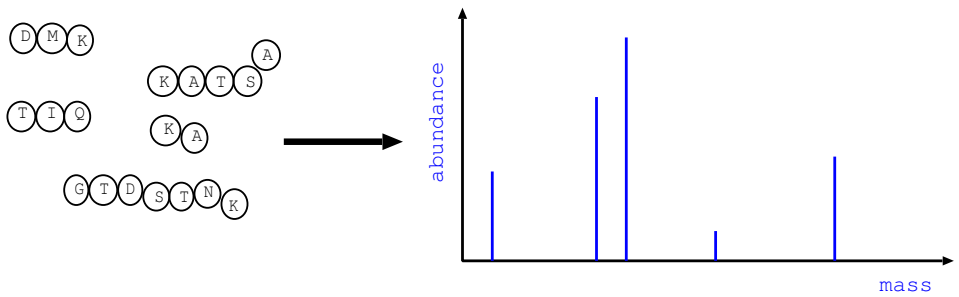
Problem

Solely measuring the mass of a protein is not sufficient for identification.



Idea

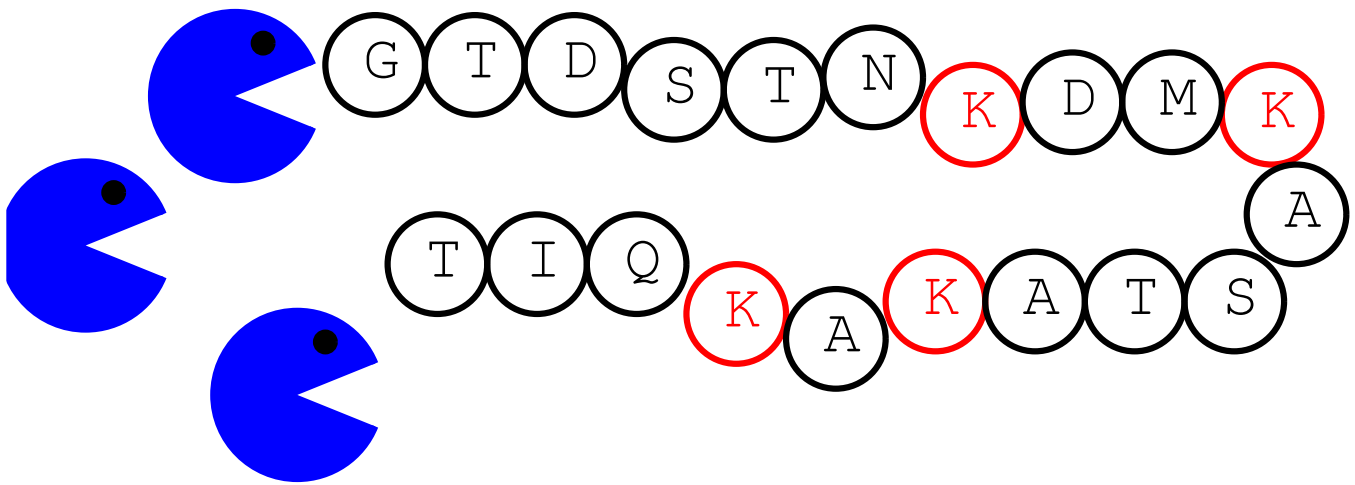
Break up the protein into smaller pieces in a deterministic way.
The spectrum of these pieces is called a *fingerprint* of the protein.



Peptide Mass Fingerprints

Enzymatic cleavage example

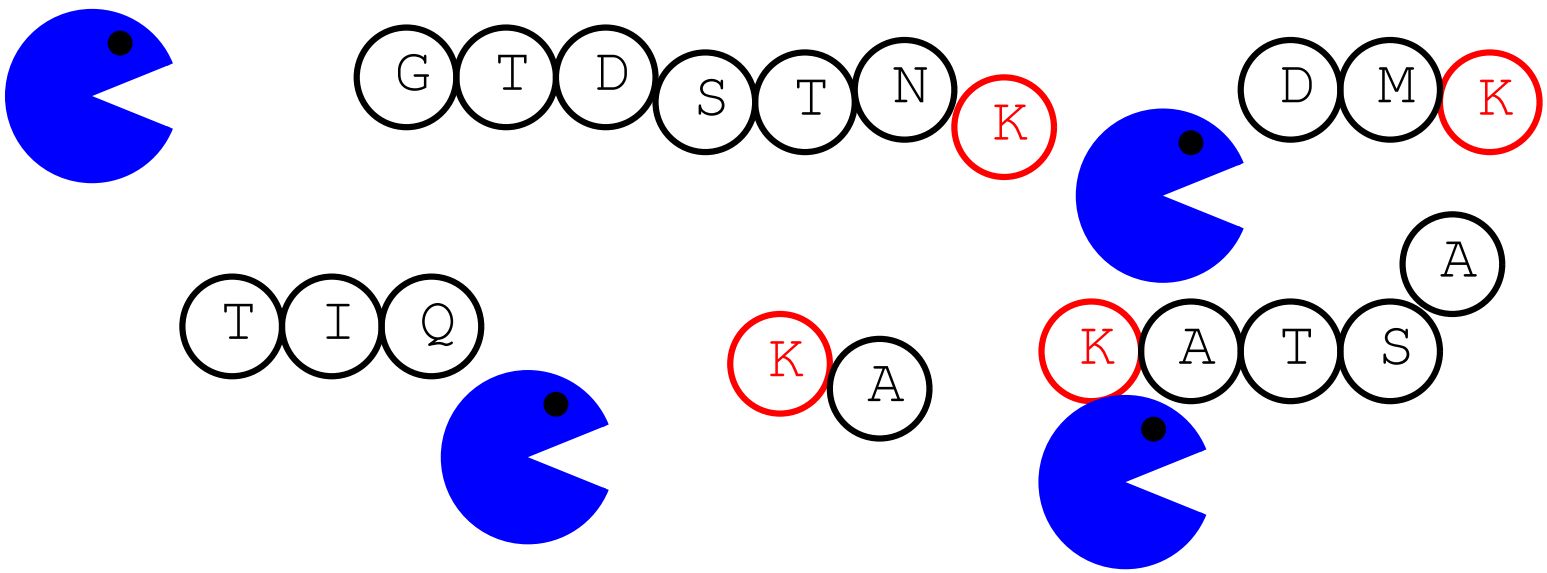
An enzyme cuts amino acid sequence after each letter **K**.



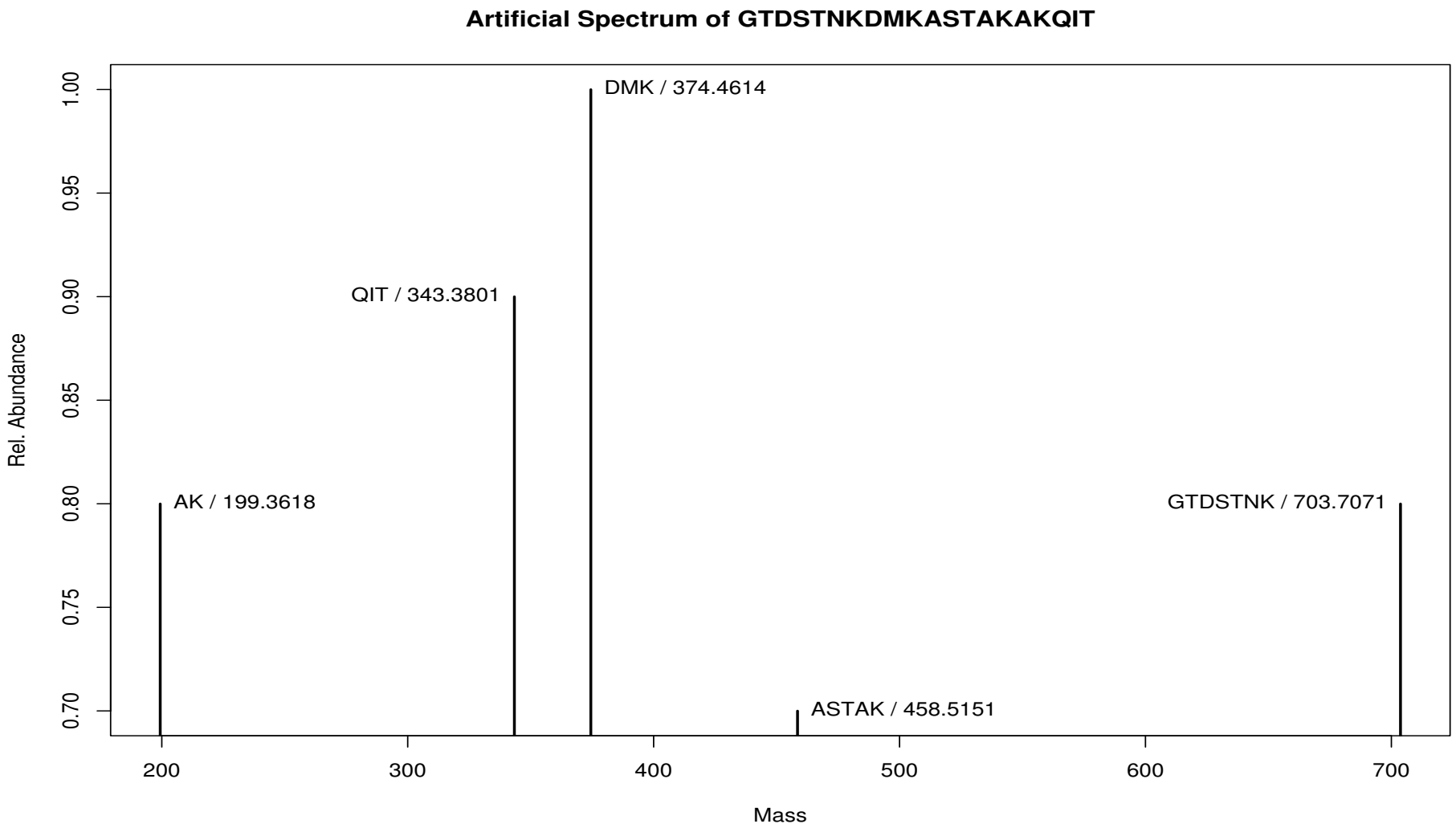
Peptide Mass Fingerprints

Enzymatic cleavage example

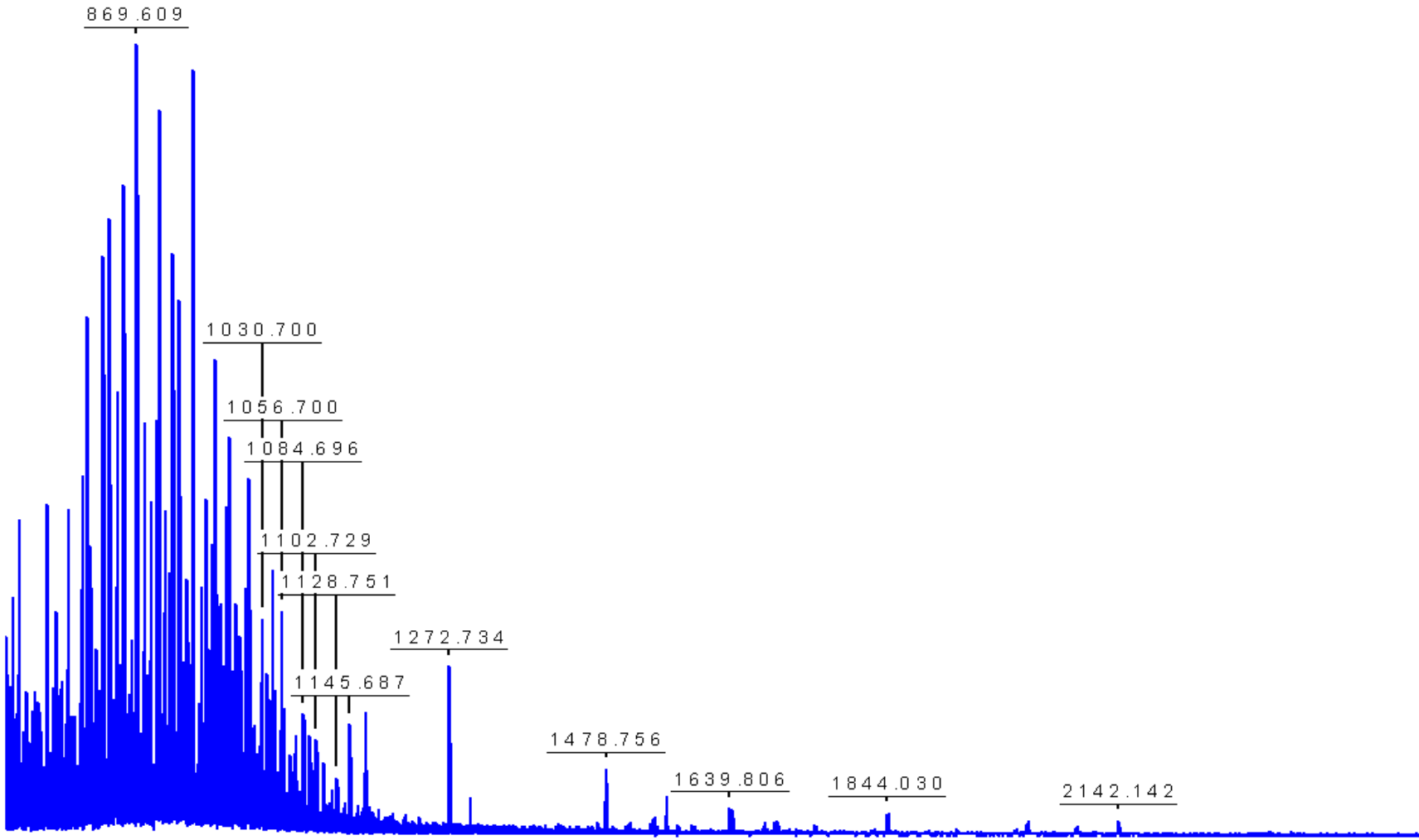
An enzyme cuts amino acid sequence after each letter **K**.



Peptide Mass Fingerprints



Real Mass Spectrum (PMF peaks annotated)



Processing the spectrum

Peak extraction

Spectra are summarized into *peak lists*, but extracting peaks is inherently difficult.

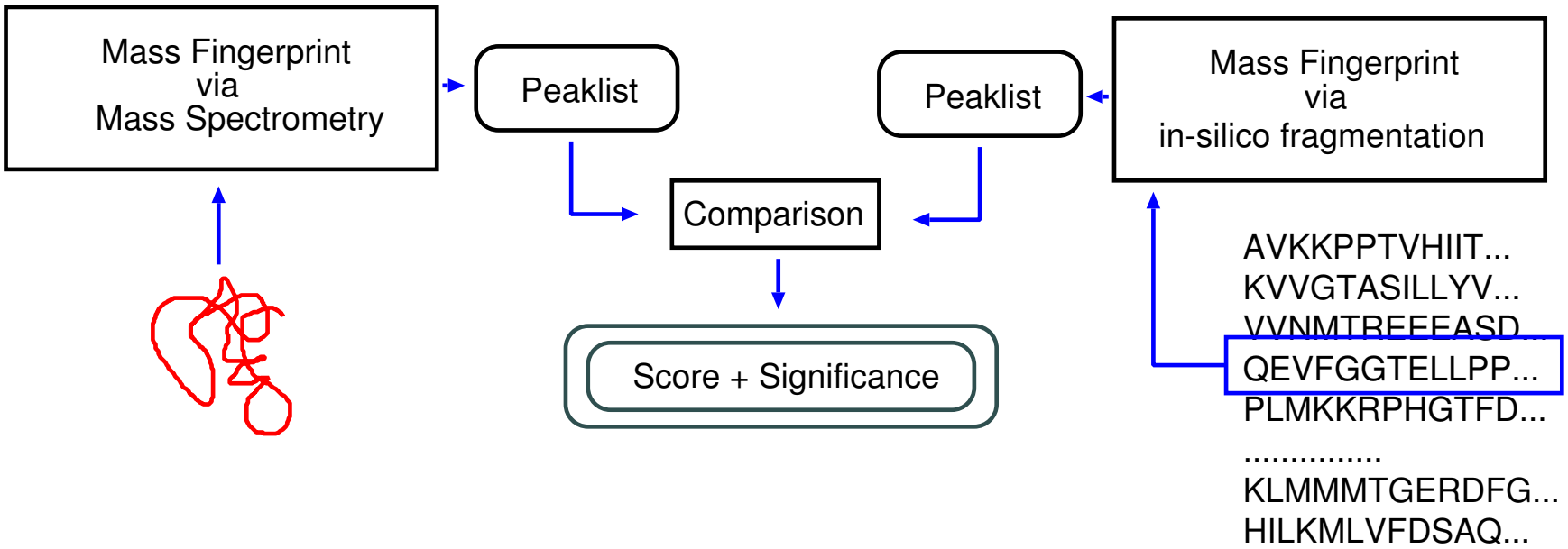
Problem: Peak lists are never correct

- Inaccurate calibration
- Probe contamination
- Peak detection
- ...

Identification

Protein Identification w/ PMF

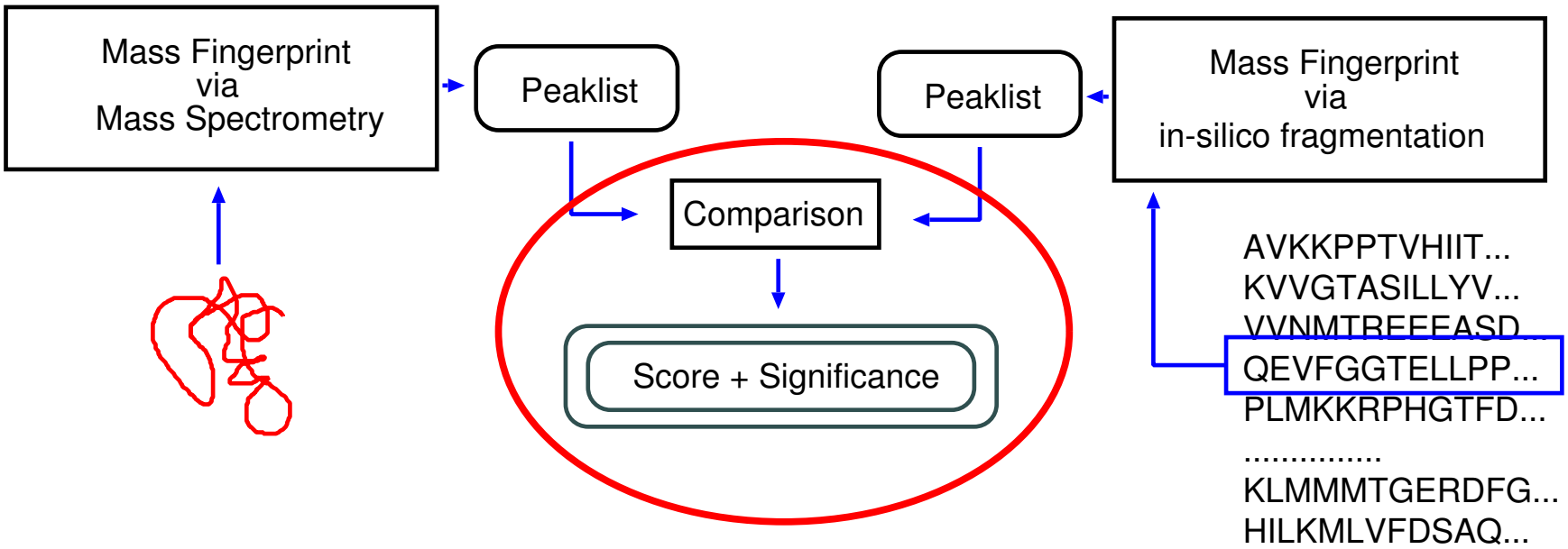
- Isolate many copies of ONE protein
- Digest it into specific smaller fragments (Mass Fingerprint)
- Make a mass spectrum of these fragments
- Compare spectrum to all predicted mass spectra from DB



Identification

Protein Identification w/ PMF

- Isolate many copies of ONE protein
- Digest it into specific smaller fragments (Mass Fingerprint)
- Make a mass spectrum of these fragments
- Compare spectrum to all predicted mass spectra from DB



Comparing Two Peak Lists

Peaklists and Empty Peaks

Let $\mathcal{S}_m$, $\mathcal{S}_p$ be an extracted and a predicted peaklist. Let ε denote a special *gap* peak.

Scoring Scheme

Each assignment between the two peak lists can be scored:

$$\begin{aligned} \text{score}(\mathcal{S}_p, \mathcal{S}_m) \quad = \quad & \sum_{\text{matched } i,j} \text{score}(i,j) \quad \text{matched peaks} \\ & + \sum_{\text{missing}} \text{score}(i,\varepsilon) \quad \text{missing peaks} \\ & + \sum_{\text{additional}} \text{score}(\varepsilon,j) \quad \text{additional peaks} \end{aligned}$$

Matching peaklists

Matching

- One-to-one peak matching
- Peak matchings should not cross
- Any peak must be matched either to a peak or to the gap peak
- Matching score mainly based on mass difference but can include other features

Best matching

Using such scoring schemes, the best peaklist matching can be computed using *standard global alignment*.

Scoring scheme example: Peak counting

Peak counting score

$$\begin{aligned} \text{score}(i,j) &= \begin{cases} 1 & |\text{mass}(i) - \text{mass}(j)| \leq \delta \\ 0 & \text{else} \end{cases} \\ \text{score}(i,\varepsilon) &= \text{score}(\varepsilon,j) = 0 \end{aligned}$$

$\delta = 10$, $\mathcal{S}_m = \{1000, 1230, 1500\}$ and $\mathcal{S}_p = \{1000, 1235, 1700\}$

Alignment

$\mathcal{S}_p$	1000	1235	ε	1700
$\mathcal{S}_m$	1000	1230	1500	ε

$\text{score}(\mathcal{S}_m, \mathcal{S}_p) = (1+1) + 0 + 0 = 2.$

Estimating the score distribution

Problem

The score distribution depends on

- Measured spectrum
- Sequence length
- Mass and probability of characters

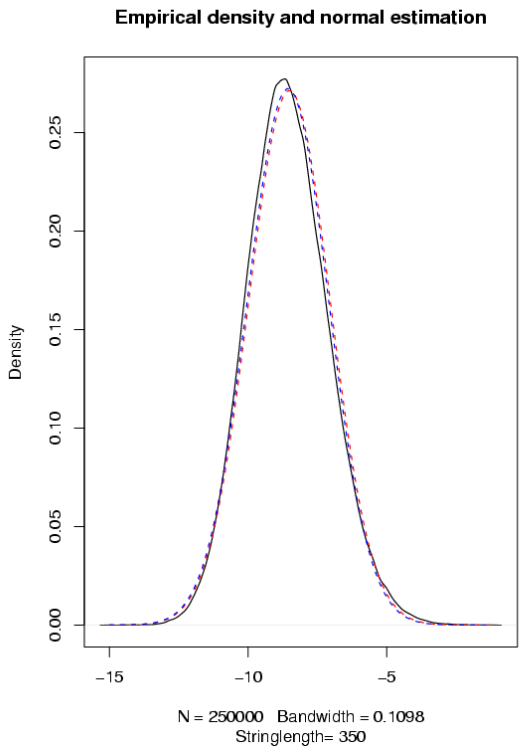
Estimation techniques

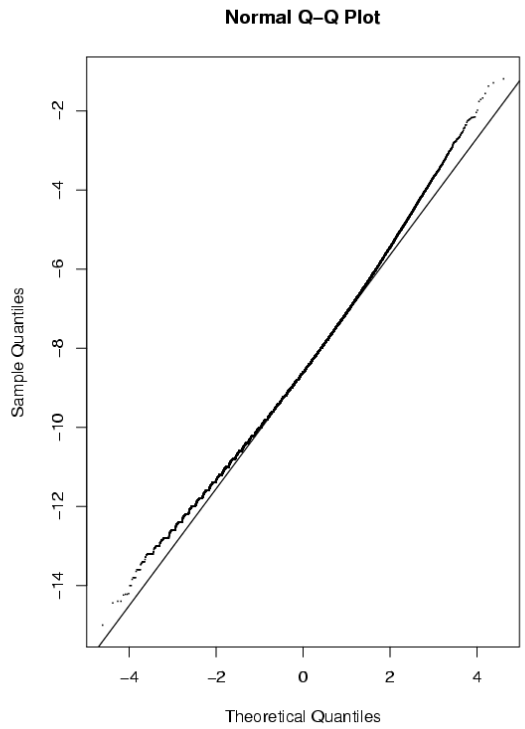
- Different null-models:
  Sampling against spectra or sampling against sequences
- Sampling against sequences
  Random or DB sequences both take long time
- Estimation of moments
  Works with certain classes of distributions

Score distribution

Claim

In most useful cases, the score distribution for fixed string length can be well approximated by a *normal distribution* and is then determined by expectation and variance. Missing and additional scores are usually very small compared to matches.





Computing moments

Main Idea

Probability of a peak corresponds to probability of a fragment of same mass in peptide.

- Discretize masses by scaling and rounding
- Compute probability of fragment of length l with mass $\neq m$
- Compute probability of string of length L to have no fragment of peak mass m
- Can all be done in preprocessing
- Estimate moments
- Compute p-value

Computing moments

Main Idea

Probability of a peak corresponds to probability of a fragment of same mass in peptide.

- Discretize masses by scaling and rounding
- Compute probability of fragment of length l with mass $\neq m$
- Compute probability of string of length L to have no fragment of peak mass m
- Can all be done in preprocessing
- Estimate moments
- Compute p-value

Fragment probability

Weighted Alphabet

We call the tuple (Σ, μ) with mass function $\mu : \Sigma \to \mathbb{N}$ an (integer) *weighted alphabet*. Define $\mu(s) := \sum_{k=1}^{|s|} \mu(s_k)$.

Fragments

Let x be the cleavage character and $\Sigma_x = \Sigma \backslash \{x\}$. The number of fragments of length l with mass m is then given by

$$c[l, m] = \sum_{\sigma \in \Sigma_x, \mu(\sigma) \leq m} c[l-1, m-\mu(\sigma)]$$

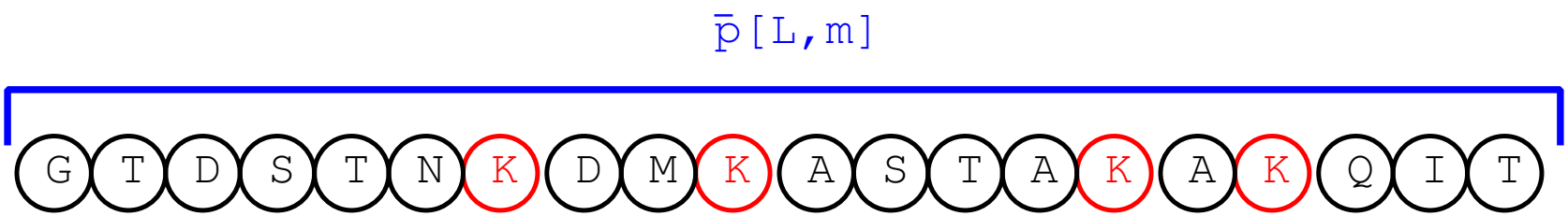
and for uniform character distribution we get the probability

$$r[l, m] = 1 - \frac{c[l, m]}{|\Sigma_x|^l}$$

Probability in Strings

Main idea

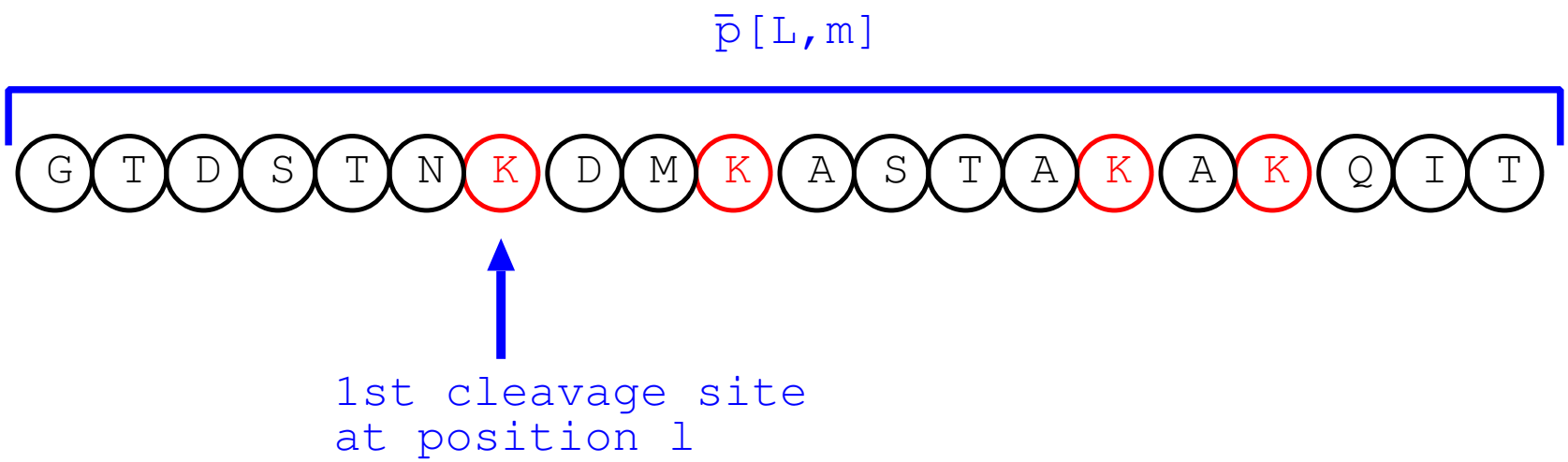
We compute prob. of string having NO fragment of mass m. Then the very first fragment must not have mass m and the following string must have no fragment of mass m. Iterate.



Probability in Strings

Main idea

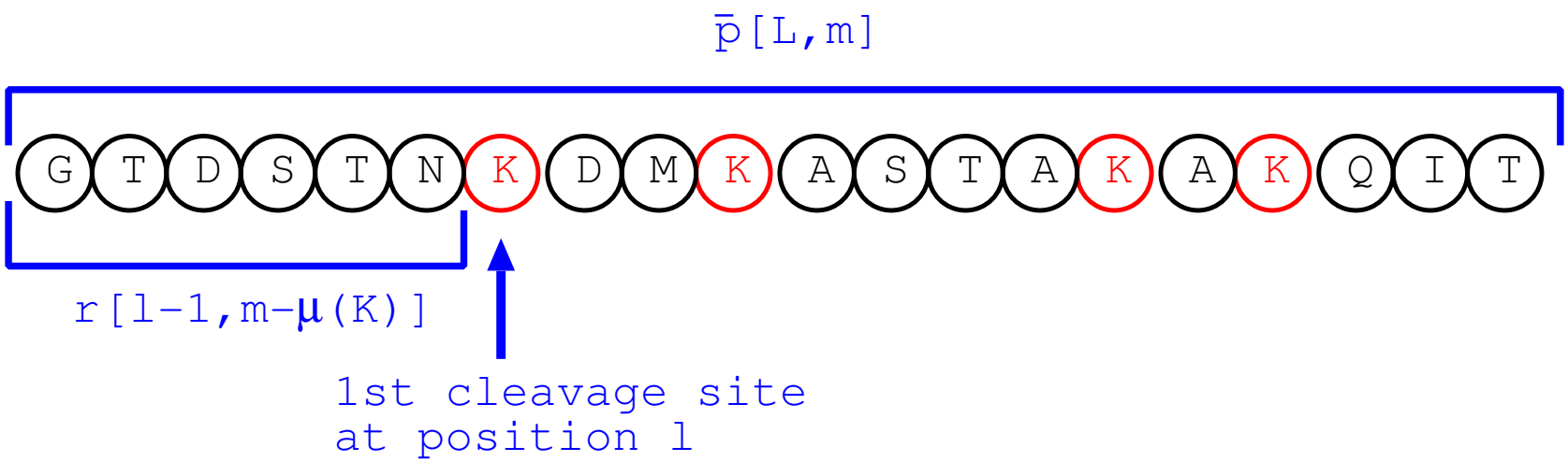
We compute prob. of string having NO fragment of mass m. Then the very first fragment must not have mass m and the following string must have no fragment of mass m. Iterate.



Probability in Strings

Main idea

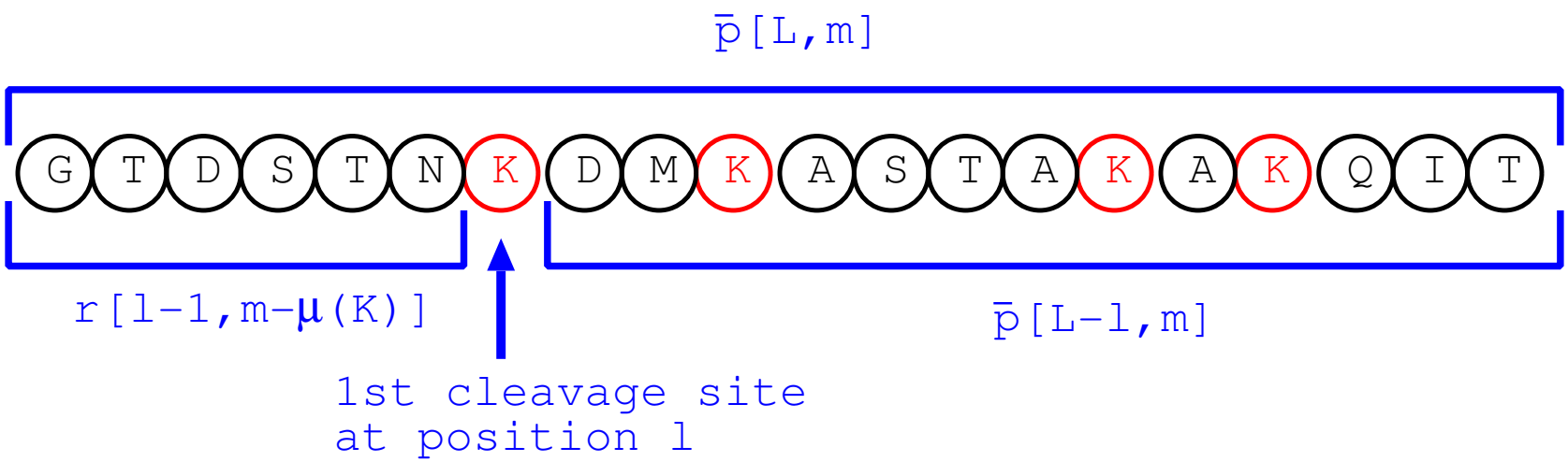
We compute prob. of string having NO fragment of mass m. Then the very first fragment must not have mass m and the following string must have no fragment of mass m. Iterate.



Probability in Strings

Main idea

We compute prob. of string having NO fragment of mass m. Then the very first fragment must not have mass m and the following string must have no fragment of mass m. Iterate.



Probability in Strings

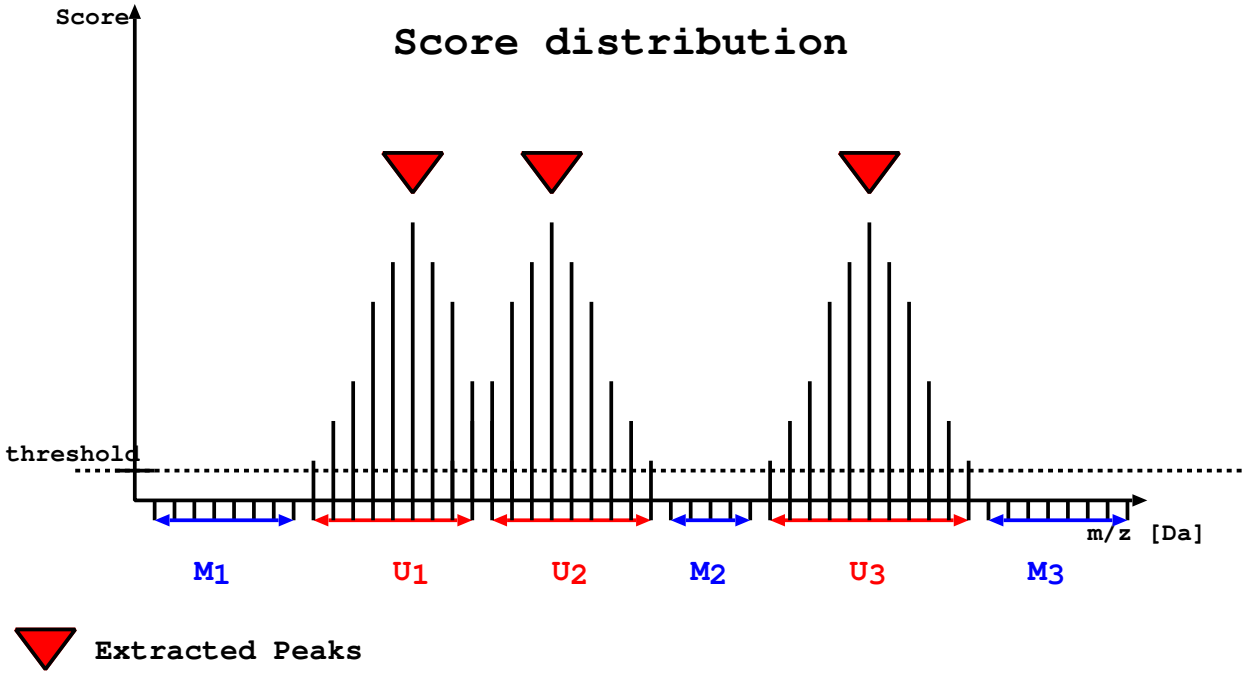
Main idea

We compute prob. of string having NO fragment of mass m. Then the very first fragment must not have mass m and the following string must have no fragment of mass m. Iterate.

The prob. of $s \in \Sigma^L$ to have NO fragment of mass m is given by

$$\bar{p}[L,m] = r[L,m] \times \mathbb{P}\,(\text{no cleavage at all})$$
$$+ \sum_{l=1}^{L} \underbrace{r[l-1, m-\mu(x)]}_{\text{first frag.}} \times \mathbb{P}\,(\text{first cleavage at } l) \times \underbrace{\bar{p}[L-l,m]}_{\text{suffix left}}$$

Expected match score of a peak



The expected value of extracted peak j with support $\mathcal{U}_j$ is

$$\mathbb{E}(\text{matchscore(j)}) = \sum_{m \in \mathcal{U}_j} p[L, m] \times \text{score}(\text{mass(j)}, m)$$

Conclusion

Main features

- Scoring schemes allow very flexible identification routines
- Computation of significance is database independent
- Extension to other cleavage schemes possible
- Extension to nonuniform alphabets and to isotope masses straightforward