[redacted]

SIRET Research Group
Department of Software Engineering
Faculty of Mathematics and Physics
Charles University in Prague
Czech Republic

Improving the Similarity Search of Tandem Mass Spectra using Metric Access Methods

Jiří Novák, Tomáš Skopal, David Hoksza and Jakub Lokoč

Program of Presentation

- Introduction
- Tandem Mass Spectrometry (MS/MS)
  - basic principles
  - existing methods for interpretation of the mass spectra
  - common problems of interpretation
- Similarity Search Approaches
  - angle distance (cosine similarity)
  - parametrised Hausdorff distance
  - TriGen
- Experiments
- Conclusions and Future Work

Introduction

- biological motivation
  - all organisms – DNA – proteins

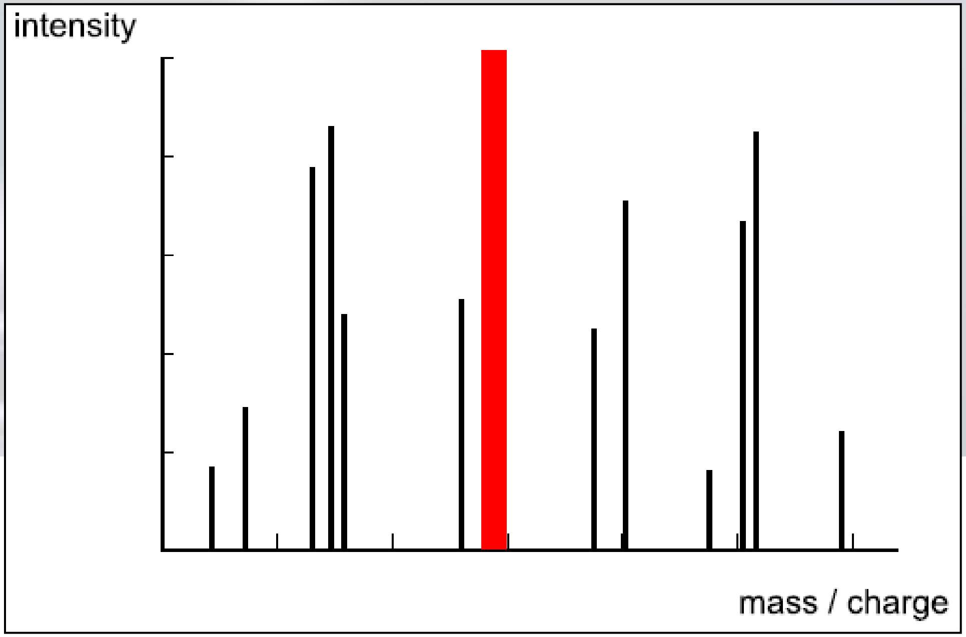
- proteins
  - cells function and structure
  - basic blocks – amino acids
  - linear sequence of amino acids
    (”linear sequence over 20-letter subset of the English alphabet”)

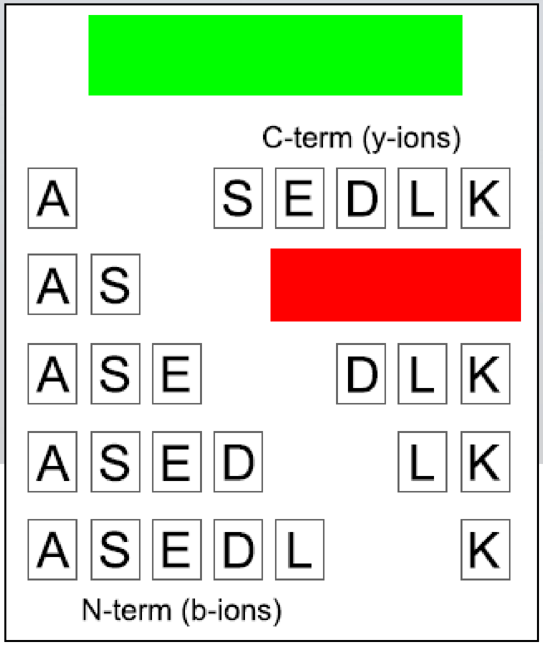
- peptides
  - short sequences

Tandem Mass Spectrometry (MS/MS)

- method for unknown protein sequences identification
  - proteins are splitted to peptides (one spectrum for each peptide is captured)
  - peptides are splitted to fragments
  - mass to charge ratio (x axis); intensity of occurrence (y axis)
  - y-ions (“from the right”); b-ions (“from the left”)

MGLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLE
KFDKFKHLKSEDEMKASEDLK...

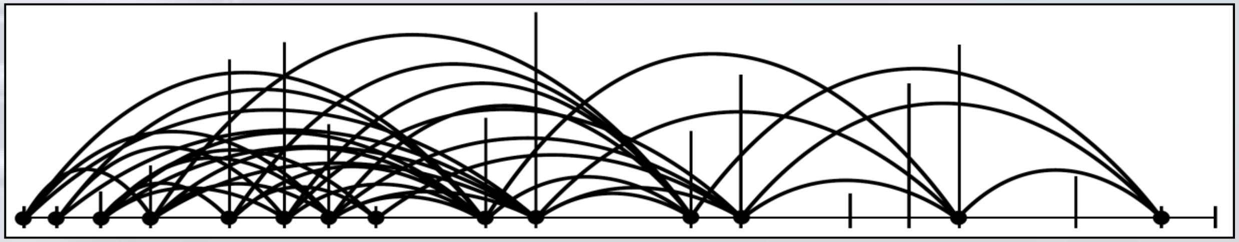




Interpretation of Spectra

- main idea: different amino acids ~ different masses

- graph approach "de novo"
  - direct spectra interpretation using graph algorithms
  - many paths in graph represent many peptide sequences corresponding to an experimental spectrum; quality of identification is about 30%



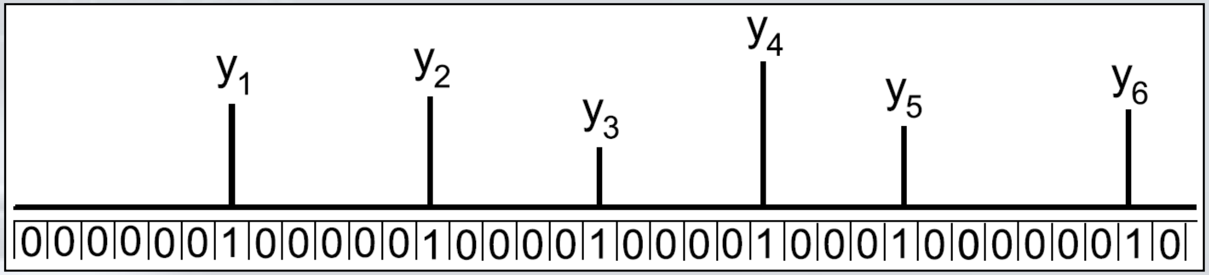
- database approach
  - search database of already known protein sequences
  - theoretical spectra are generated from stored sequences and compared with experimental spectra

Typical Problems of Interpretation

- noise
  - up to 80% of peaks
  - peaks of fragment ions with unpredictable chemical structure
- single amino acids (or groups) with similar masses can be mistaken
- some peaks important for identification (y or b-ions) are missing
  - fragment ions do not arise
- modifications of amino acids
  - amino acids masses are changed

Angle Distance (d_A)

- cosine similarity approaches are commonly mentioned in literature
- high-dimensional boolean vectors; compact representation <7, 13, 18, 23, 27, 34>
- bad indexability

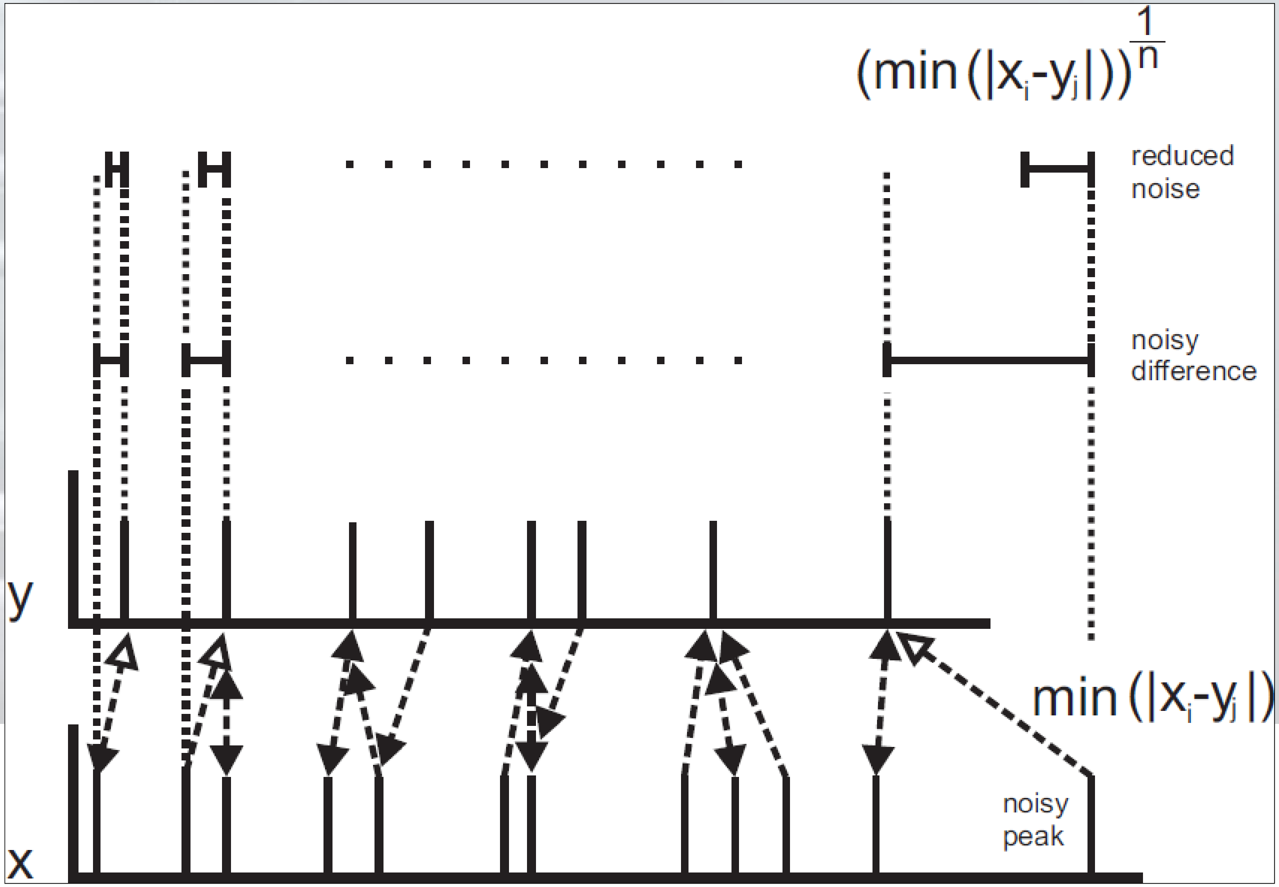


0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	1	0

- precursor mass
  - mass of a peptide before splitting (known as an additional information)
- precursor mass filter
  - spectra are indexed by their precursor mass
- $d'_A = d_A$ + precursor mass filter
  - indexable very well
  - it supports only spectra without chemical modifications

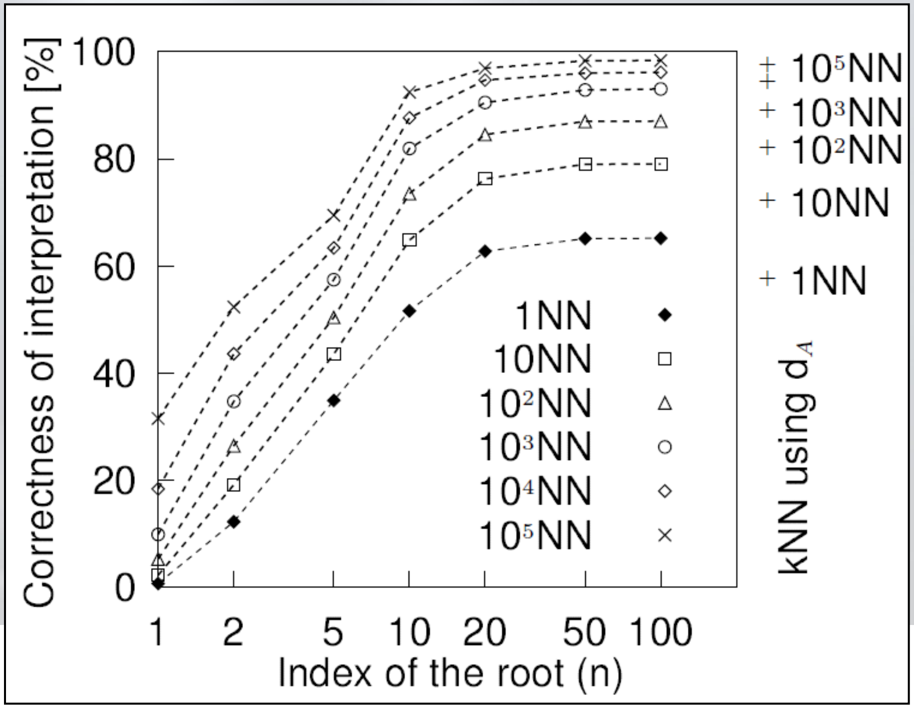
Parametrised Hausdorff Distance (d_{HP})

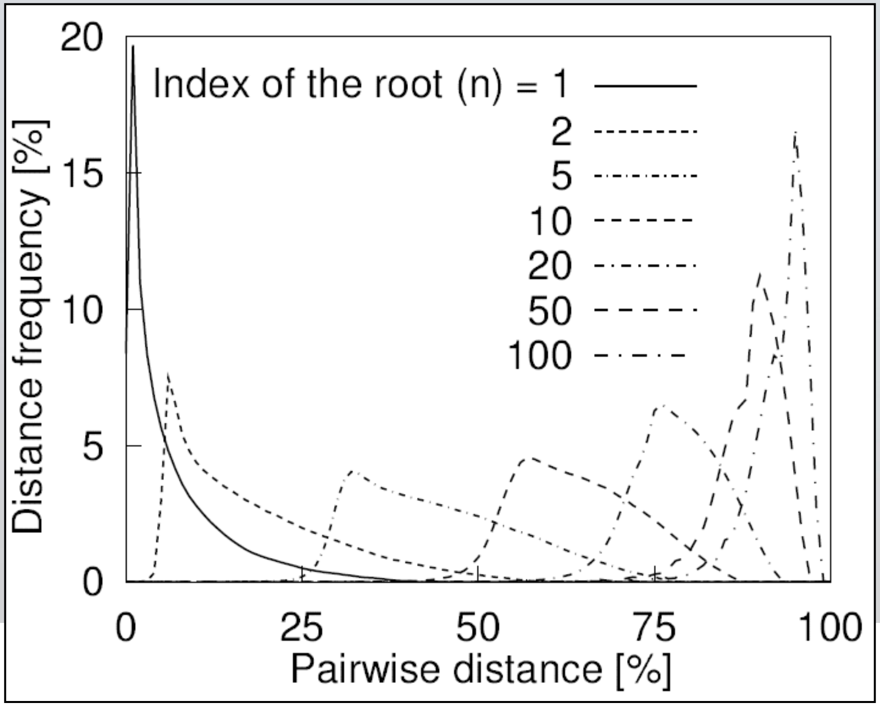
- for each number in the compact representation, the number with minimum difference in the other vector is found
- the average of n^th roots from the set of minima is computed
- d_{HP} can be also combined with precursor mass filter (for the spectra without chemical modifications)



Parametrised Hausdorff Distance (d_{HP})

- increasing n in nth root function
  - + the impact of noise peaks is lower (i.e., the similarity between the spectra is modeled better)
  - + the distance is semimetric (n ≥ 2)
  - – the indexability is worse



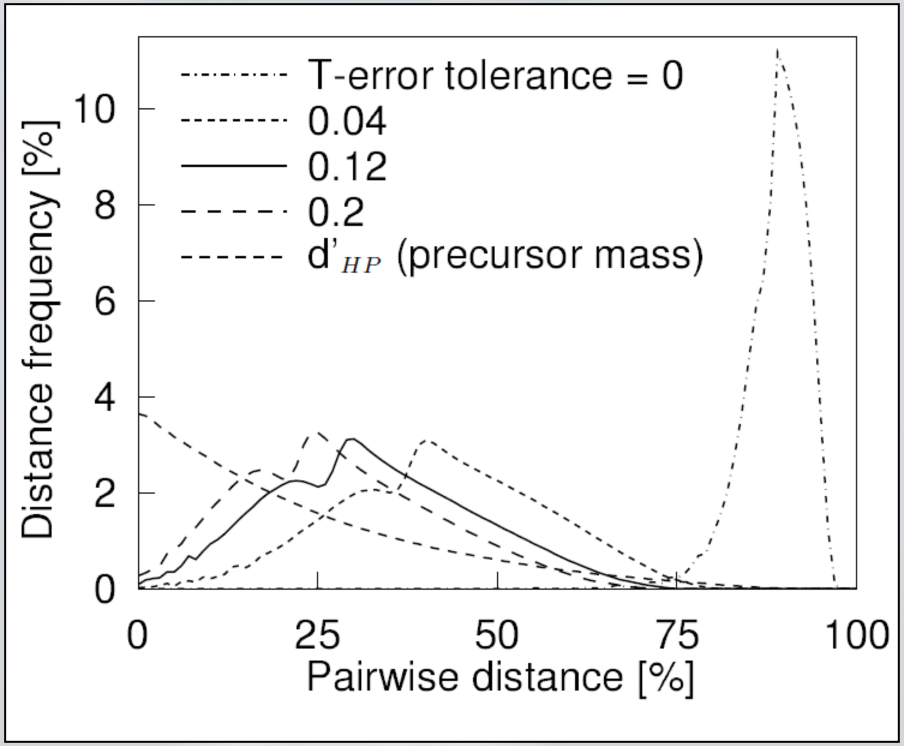


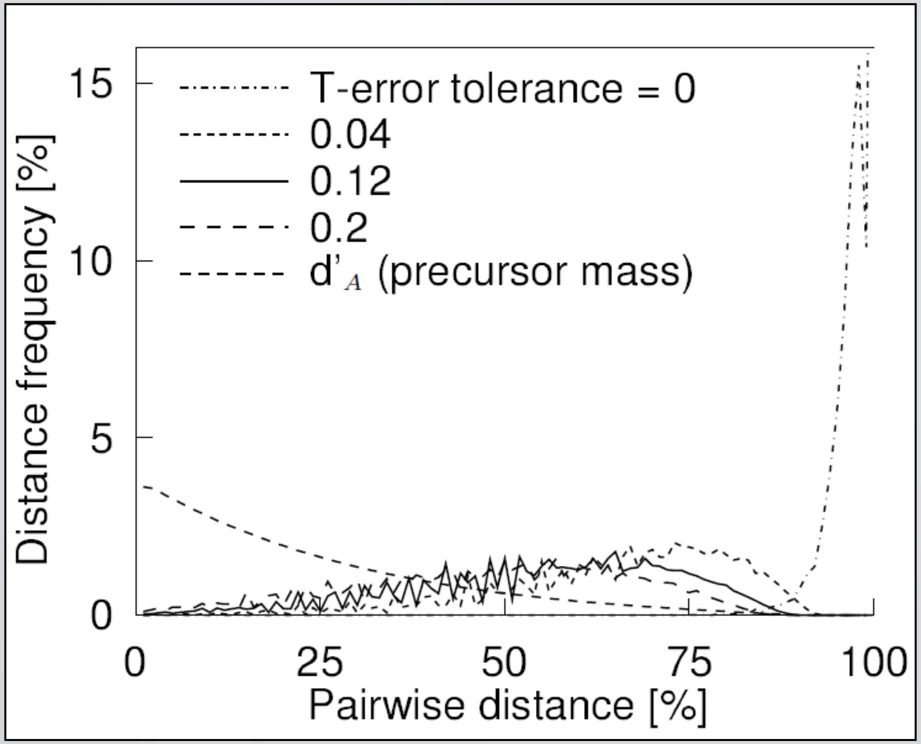
TriGen Algorithm

- controls the metricity (T-error) of the function *v*
  - the ratio of triplets, which do NOT satisfy the triangle inequality
- T-modifier
  - either concave or convex increasing function
  - e.g., Fractional-Power (FP) or Rational-Bézier-Quadratic (RBQ) modifier
  - concave function (w > 0)
    - increases the number of triplets
    - indexability is worse
    - exact search, but slower
  - convex function (w < 0)
    - decreases the number of triplets
    - indexability is better
    - approximate search, but faster
- M-tree, Pivot Table

$$\mathrm{FP}(v,w) = \begin{cases} v^{\frac{1}{1+w}} & \text{for } w > 0 \\ v^{1-w} & \text{for } w \leq 0 \end{cases}$$

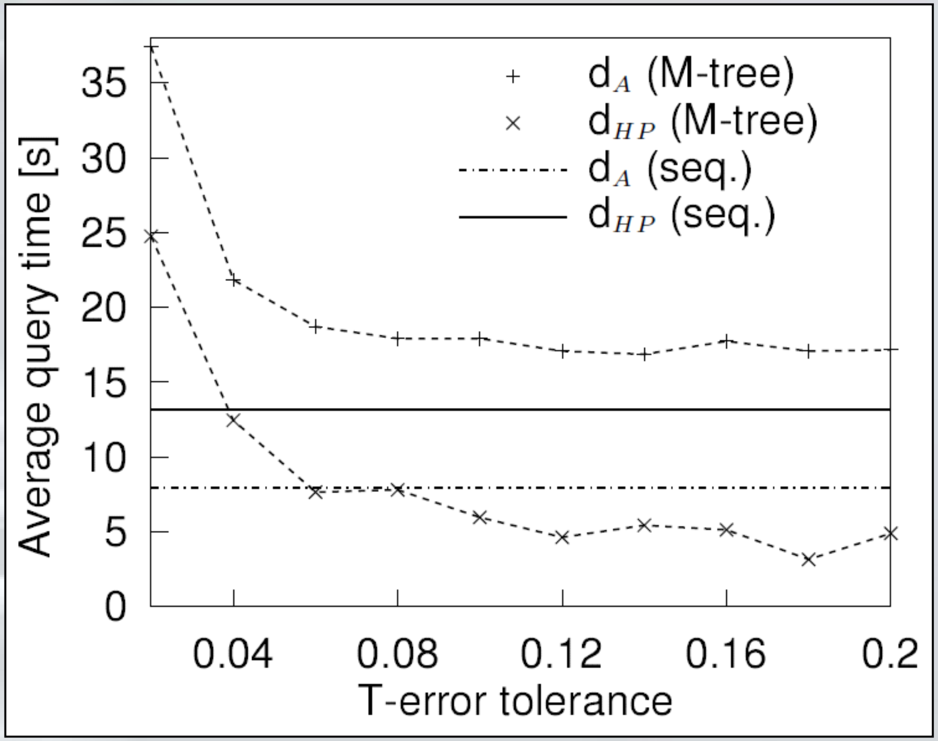
Indexability of d_{HP} and d_A





- d_{HP} – the indexability is better with increasing T-error tolerance
- d_A – about 35% of all pairwise distances in d_A=1 (uncorrectable)
- d'_{HP} and d'_A – indexable very well

Average Query Time



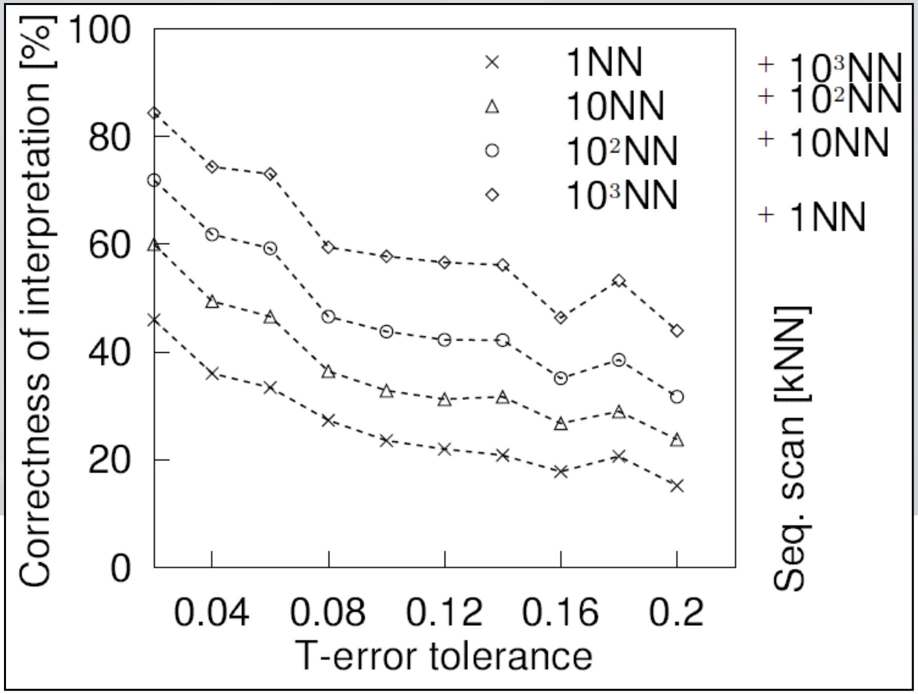
- d_{HP} – 1.6x faster than sequential scan
- d_A – 2.5x slower
- d'_{HP} and d'_A – 32.9x faster and 19.8x faster

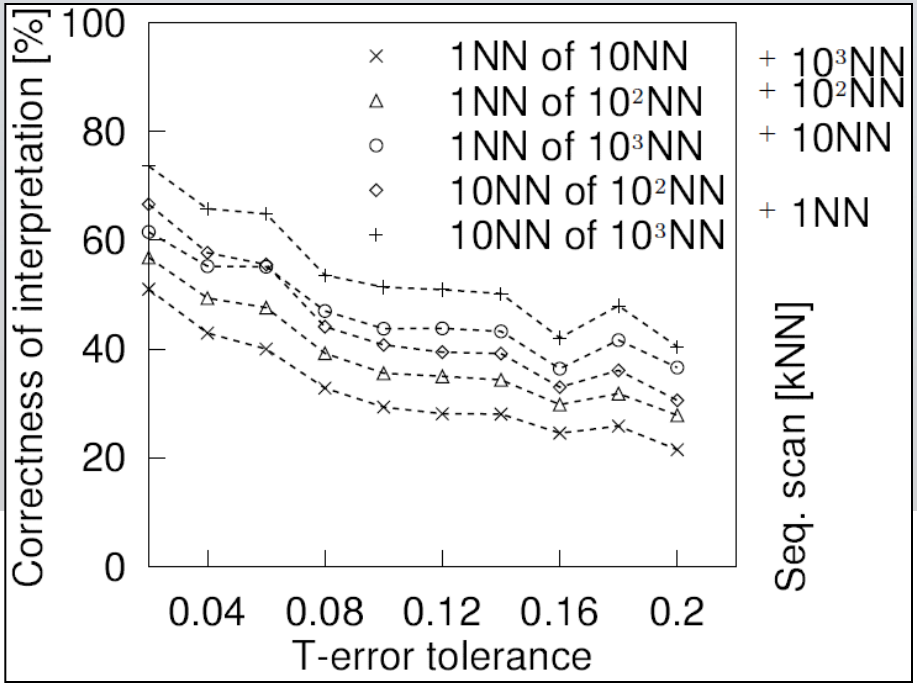
Correctness of Identification - kNN Queries

- correct peptide sequences are cumulated among a few nearest neighbors

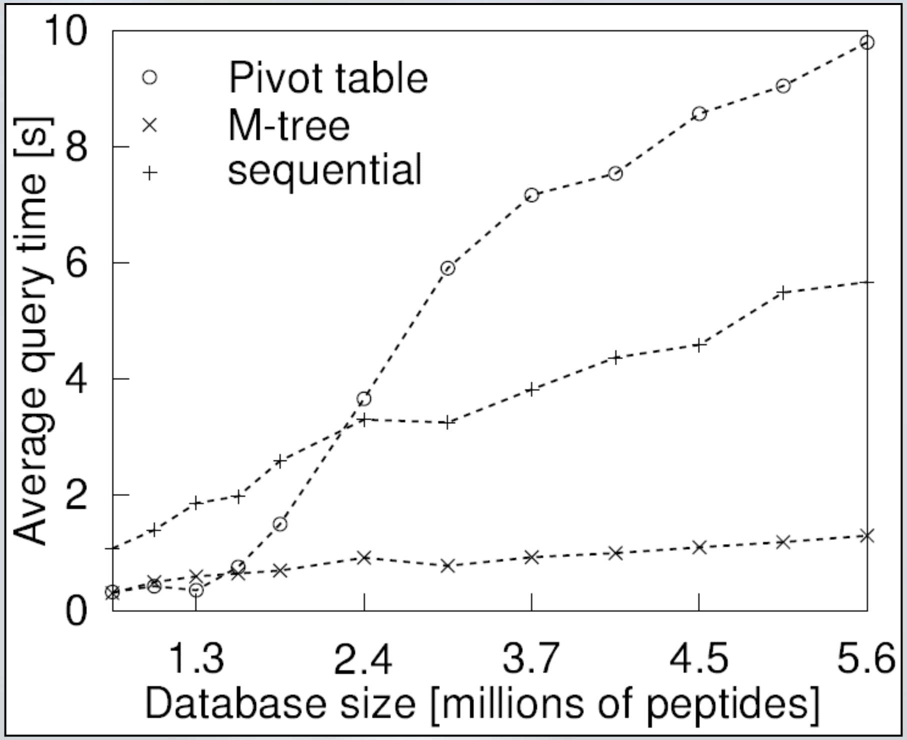
- 1-NN taken from the 100NN result is more likely to be correct than when taking 1-NN from 10NN result

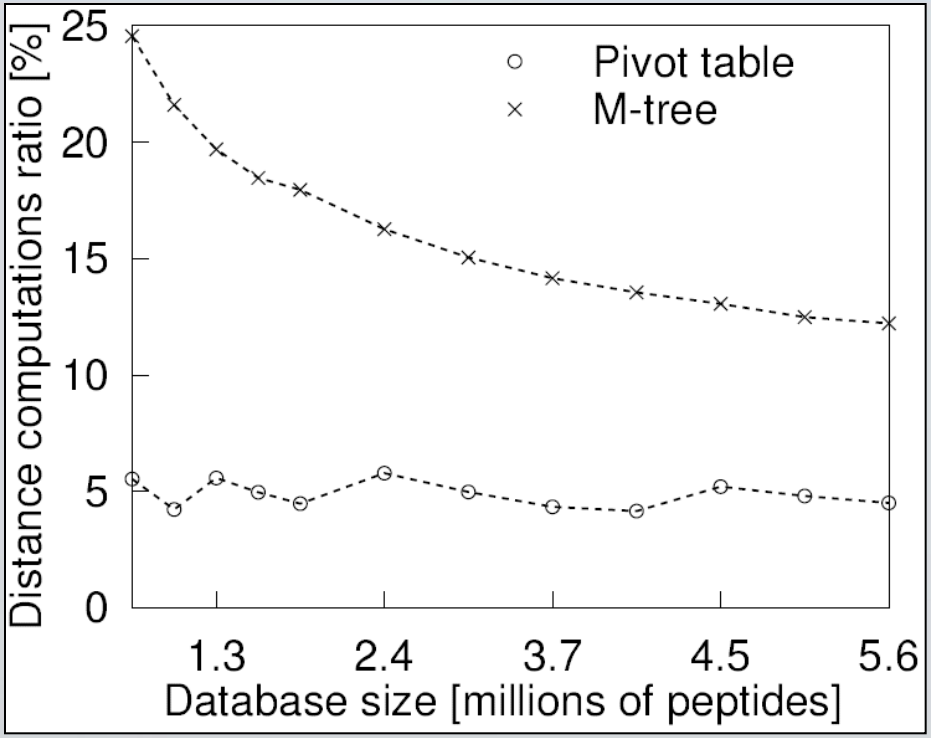
- e.g., at T-error tol. 0.06, correctness 75%, speed-up 1.7x, DC ratio 9.7%
- 1.4x higher for d_{HP} than d_A
- d'_{HP} 85.7% and d'_A 89.6%





M-tree and Pivot Table Comparison





- the Pivot table is faster than M-tree as long as all its blocks are stored in main memory, otherwise it becomes inefficient (moreover, it is outperformed by sequential scan)

- distance computations are misleading for Pivot tables

Conclusions

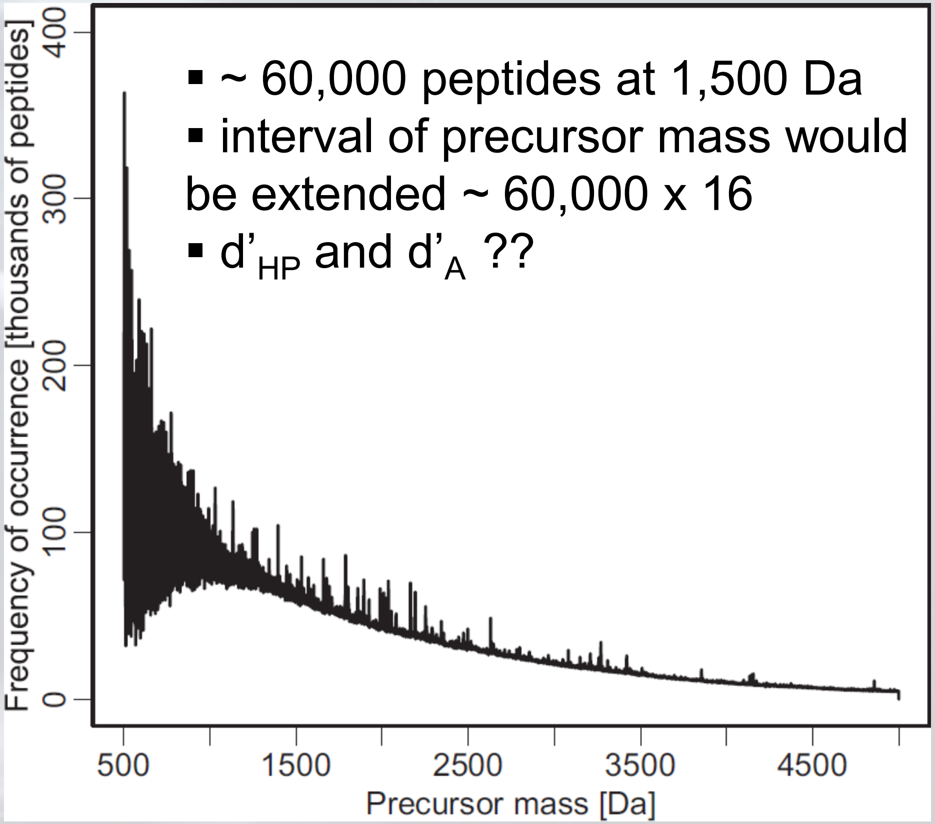
- parametrised Hausdorff distance (d_{HP})
  - models the similarity among spectra very well
  - can be utilized by MAMs when TriGen algorithm is employed
  - if the T-error is higher, then indexability is much better, the search is faster and correctness of interpretation is a little lower

- angle distance (d_A)
  - we verified that it has limitations for utilization by MAMs

- d'_{HP} or d'_A (in combination with the precursor mass filter)
  - indexable very well
  - an extension for mass spectra with chemical modifications may be very hard

Future Work

- dealing with modifications in the mass spectra - precursor mass of modified peptides can differ by more than a few tens to hundreds Daltons (e.g., M+16)



- ~ 60,000 peptides at 1,500 Da
- interval of precursor mass would be extended ~ 60,000 x 16
- d'_{HP} and d'_{A} ??

- d_{HP} seems to be suitable for particular kinds of modifications without an improvement

- NM+16INTFVPSGK
- IYFM+16AGSSK
- NSLESYAFNM+16K

- 30% correctness (1 NN)
- 50% (10NN)
- 84% (5000NN)

- PM-tree, …

Thank You…