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

Similarity Search
and
Tandem Mass Spectrometry

Jiří Novák
novak@ksi.mff.cuni.cz

Program of Presentation

- Introduction

- Tandem Mass Spectrometry (MS/MS)
  - basic principles
  - interpretation of spectra

- Similarity Search Approaches
  - angle distance (cosine similarity)
  - parametrised Hausdorff distance
  - metric access methods (MAMs)

- Experiments

- Conclusions

Introduction

- biological motivation
  - all organisms – DNA – proteins

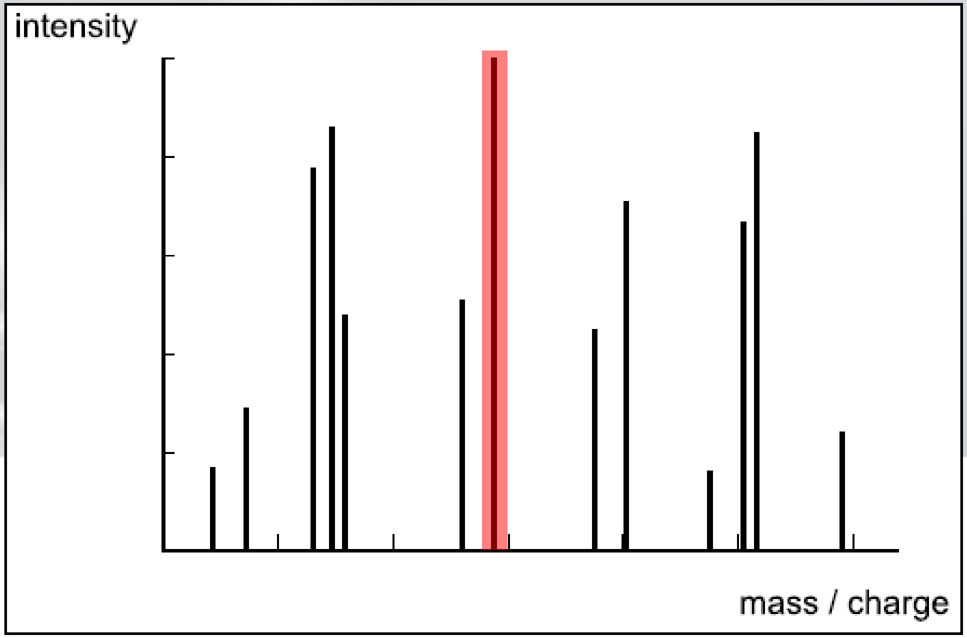
- proteins
  - cells function and structure
  - basic blocks – amino acids
  - linear sequence of amino acids
    (”linear sequence over 20-letter subset of the English alphabet”)

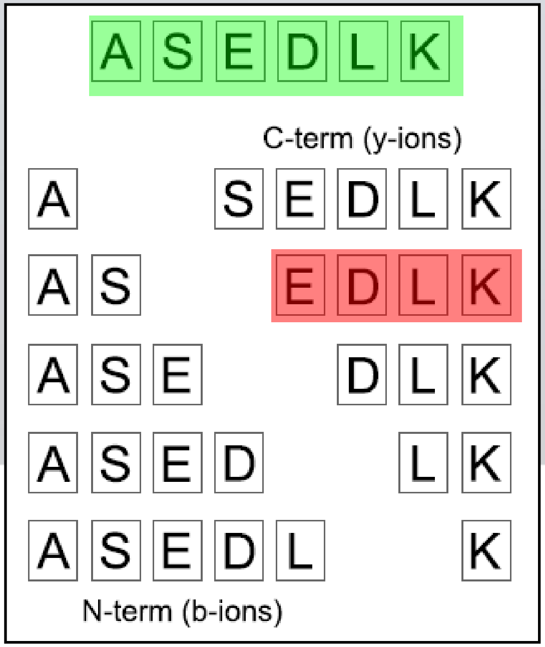
- peptides
  - short sequences

Tandem Mass Spectrometry (MS/MS)

- method for unknown protein sequences identification from an “in vitro” sample
  - 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...



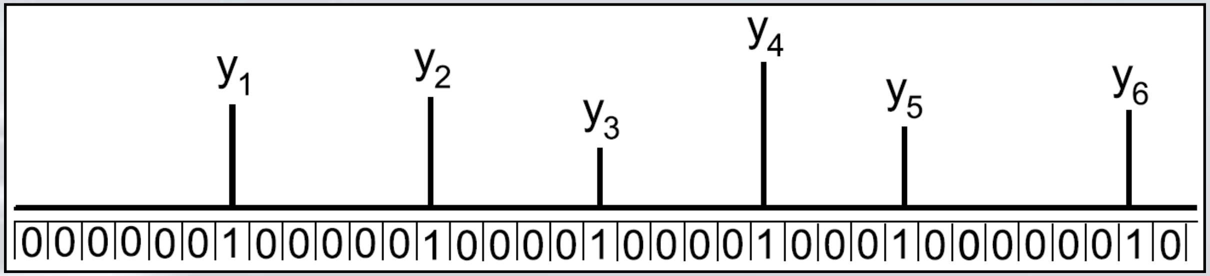


Interpretation of Spectra

- database approach
  - search database of already known protein sequences
  - theoretical spectra are generated from stored sequences and compared with experimental spectra
- typical problems
  - noise (up to 80% of peaks)
  - single amino acids (or groups) with similar masses can be mistaken
  - some peaks important for identification (y or b-ions) are missing
  - posttranslational modifications (PTMs)

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

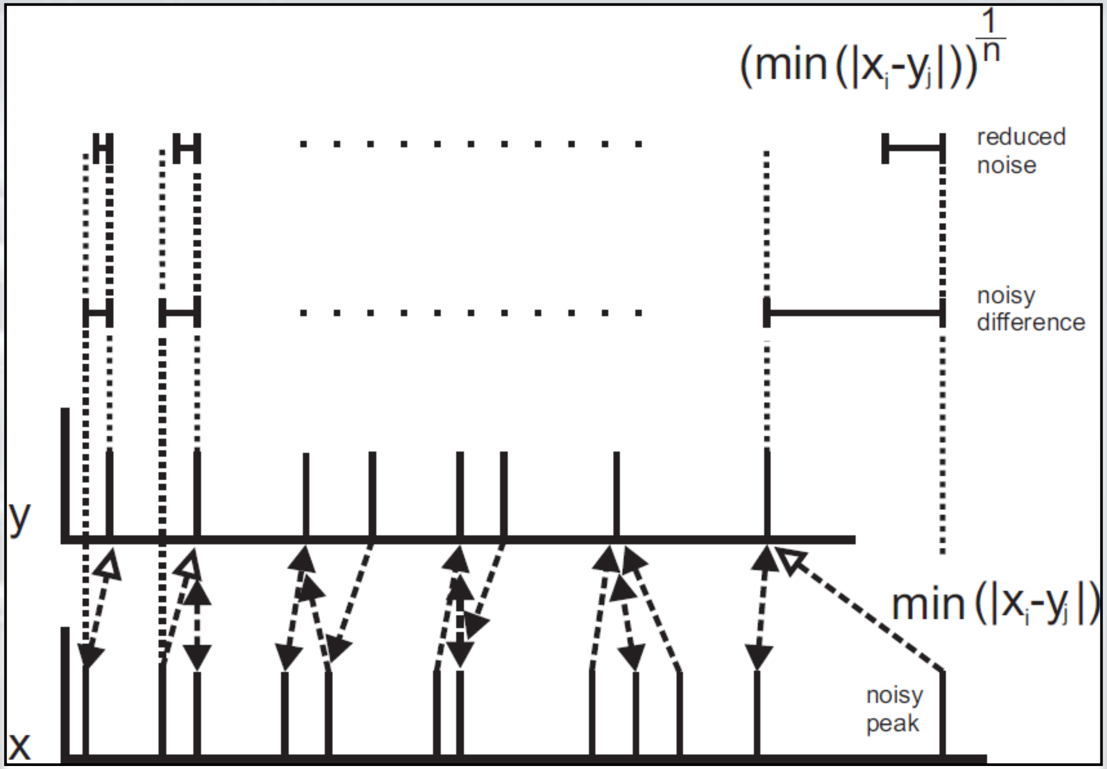


0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	1	0

$$\cos(\boldsymbol{x}, \boldsymbol{y}) = \frac{\boldsymbol{x}\boldsymbol{y}}{\|\boldsymbol{x}\| \, \|\boldsymbol{y}\|}$$

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 nth roots from the set of minima is computed



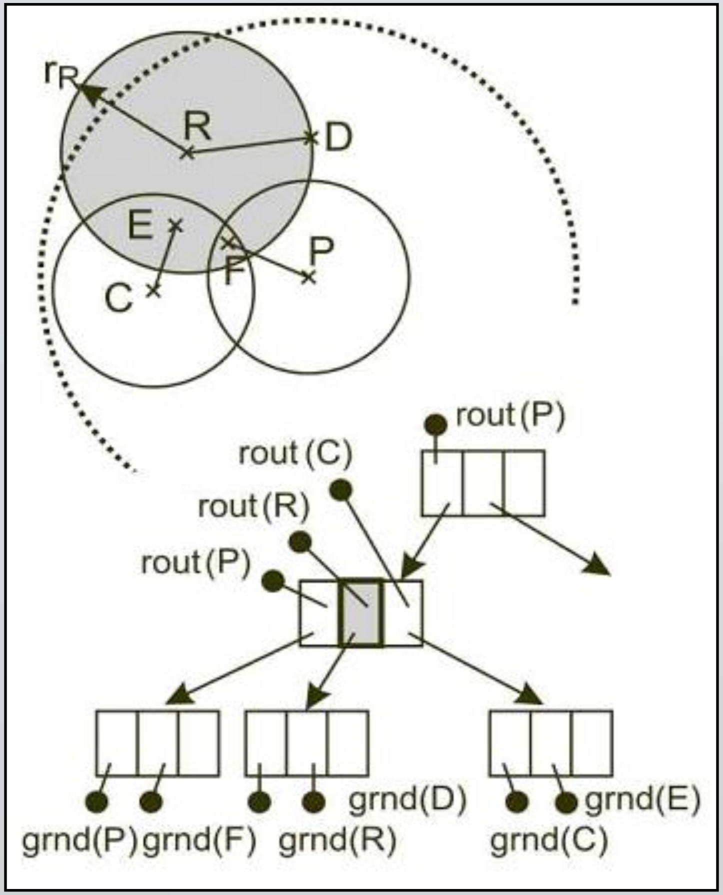
$$d_{HP}(\boldsymbol{x}, \boldsymbol{y}) = (\max(h(\boldsymbol{x}, \boldsymbol{y}), h(\boldsymbol{y}, \boldsymbol{x})))^m \qquad h(\boldsymbol{x}, \boldsymbol{y}) = \frac{\sum_{x_i \in \boldsymbol{x}} \sqrt[n]{(\min_{y_j \in \boldsymbol{y}} \{d_E(x_i, y_j)\})}}{|\boldsymbol{x}|}$$

Metric Access Methods (MAMs)

- DB index structures

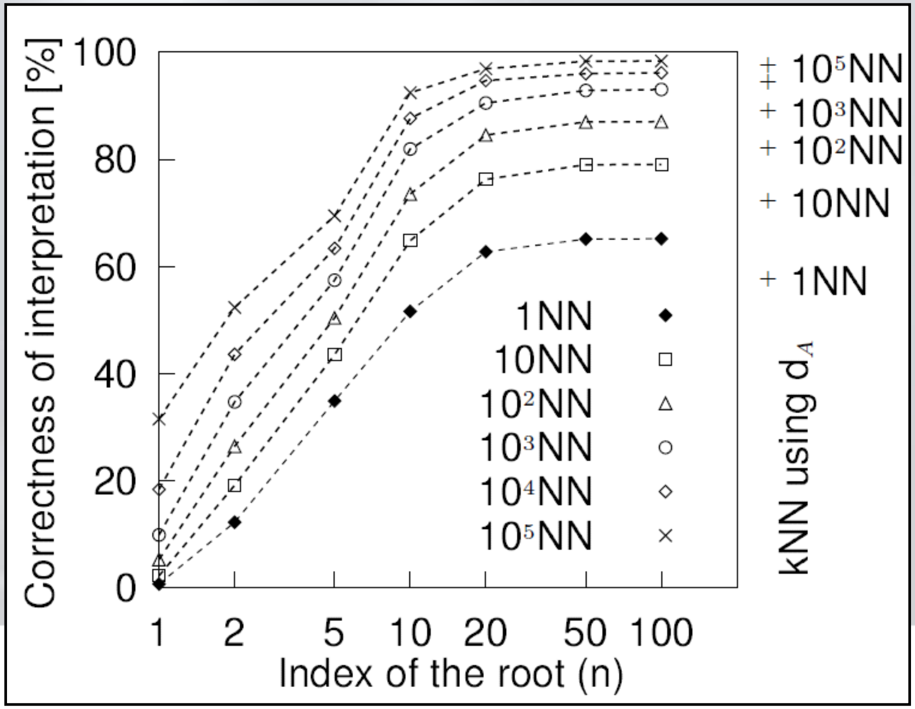
- Metric
  - qualifies the distance (or similarity) between theoretical and experimental spectra

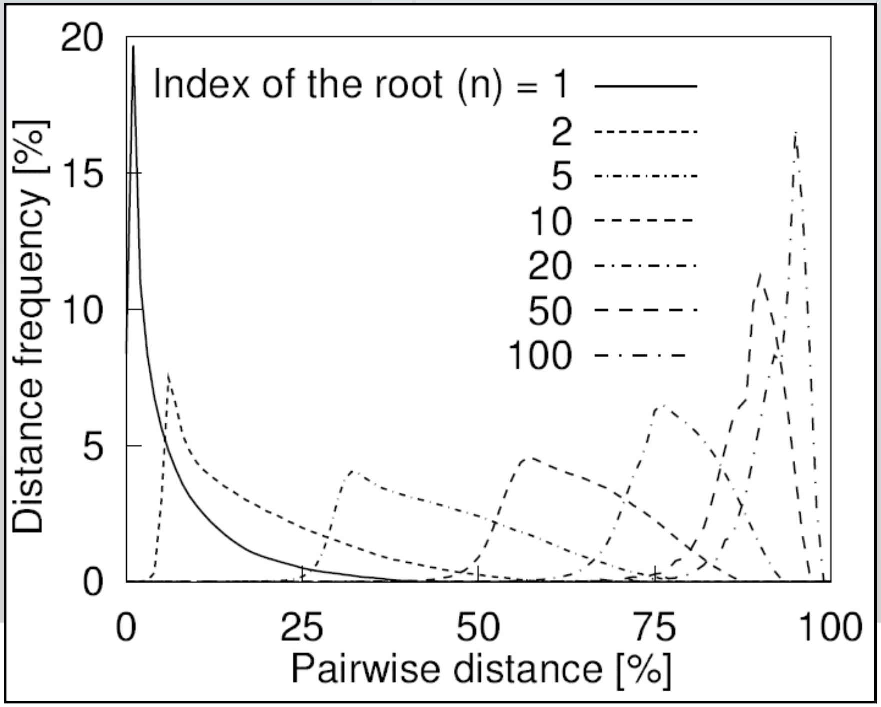
- M-tree (Metric tree)
  - dynamic and balanced tree
  - organizes objects (vectors) to n-dimensional ball regions



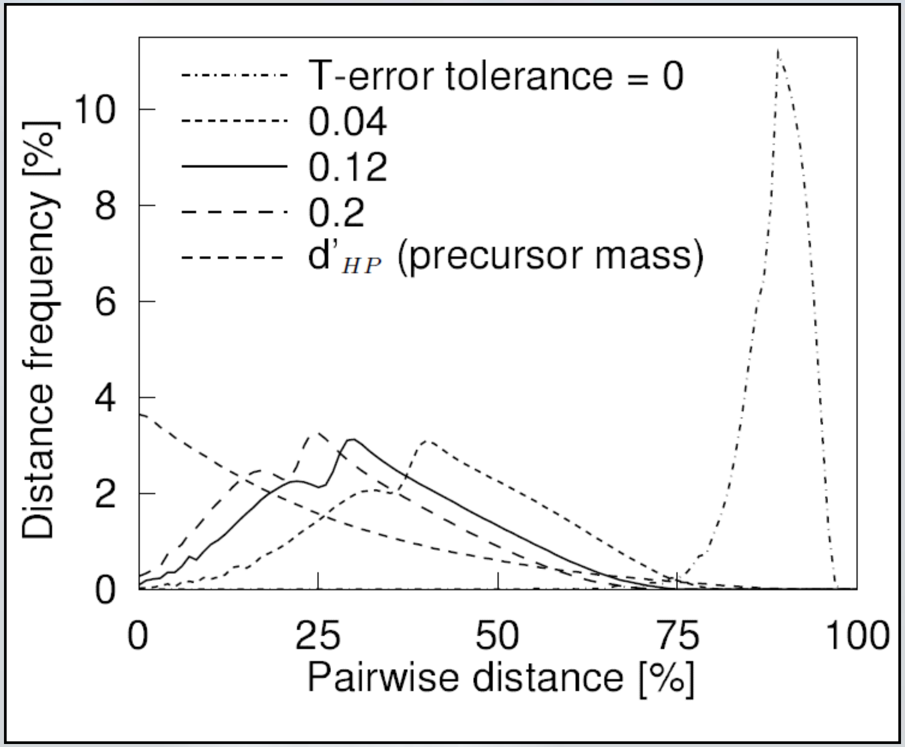
Parametrised Hausdorff Distance (d_{HP})

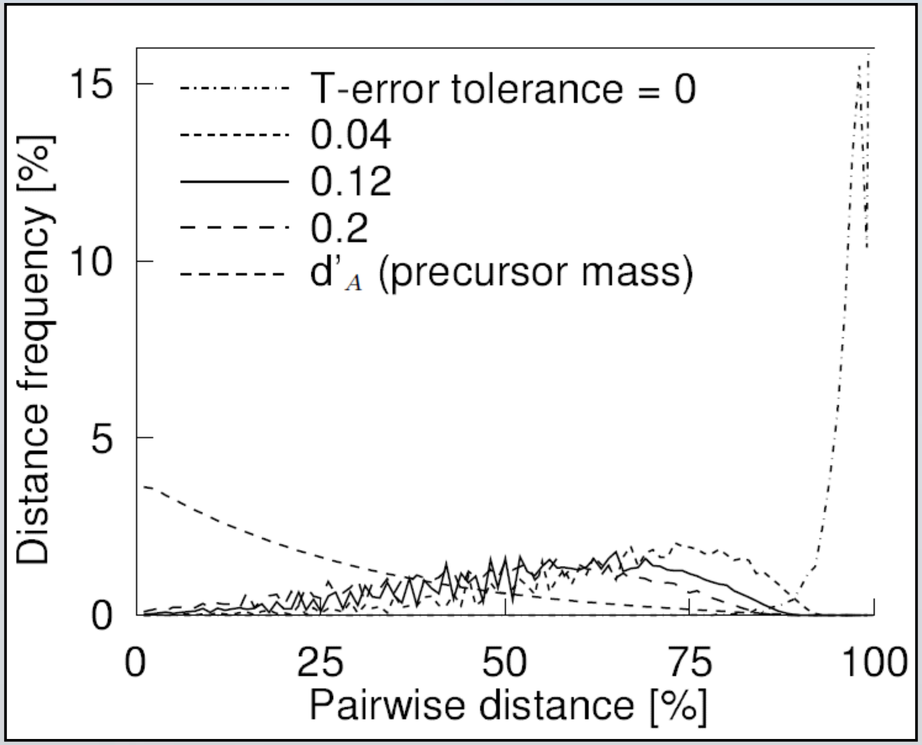
- increasing n in n^th root function
  - + the impact of noise peaks is lower (i.e., the similarity between the spectra is modeled better)
  - + the distance is semimetric (n ≥ 2)
  - – the indexability is worse





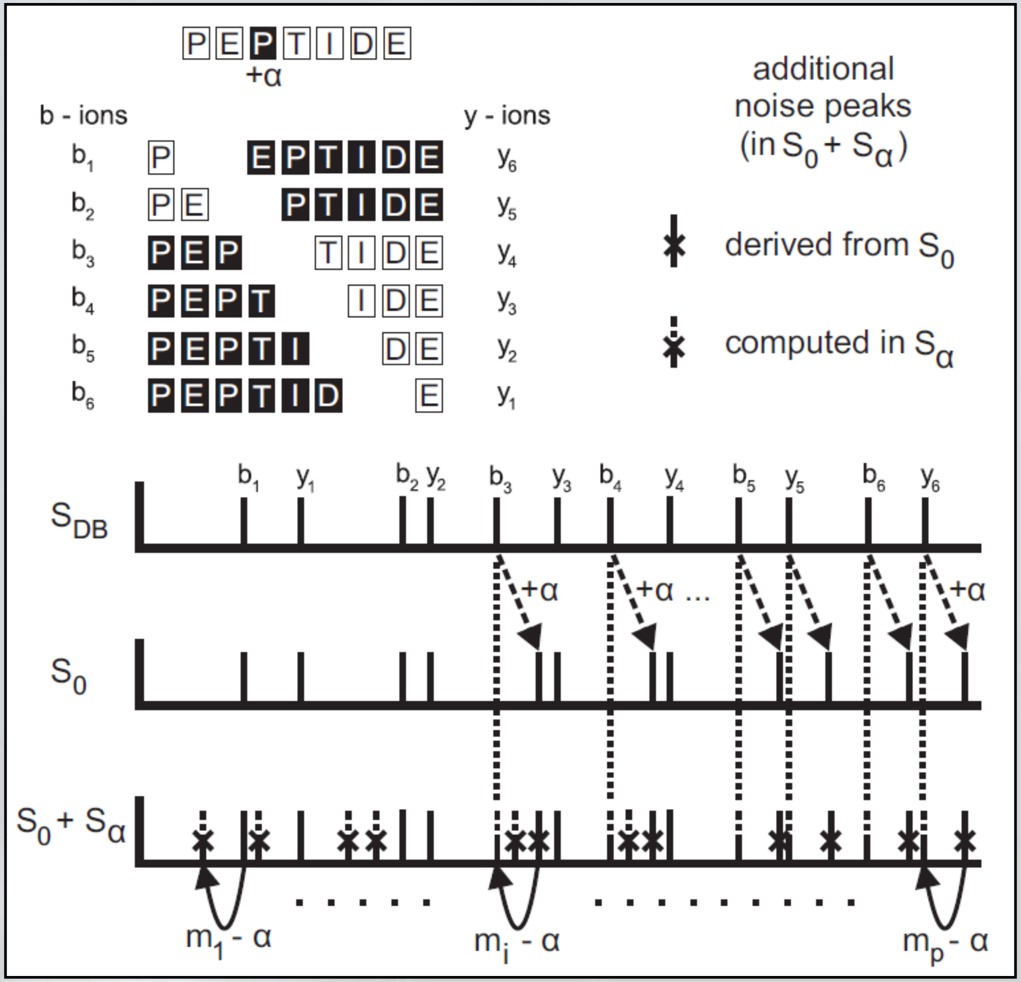
Indexability of d_{HP} and d_A

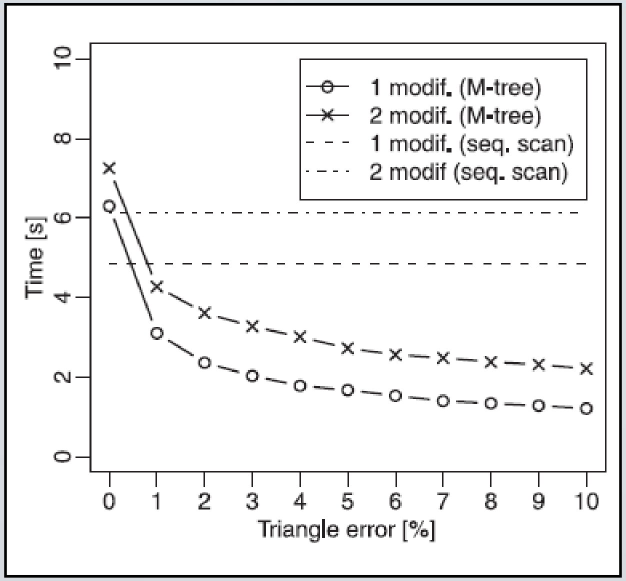


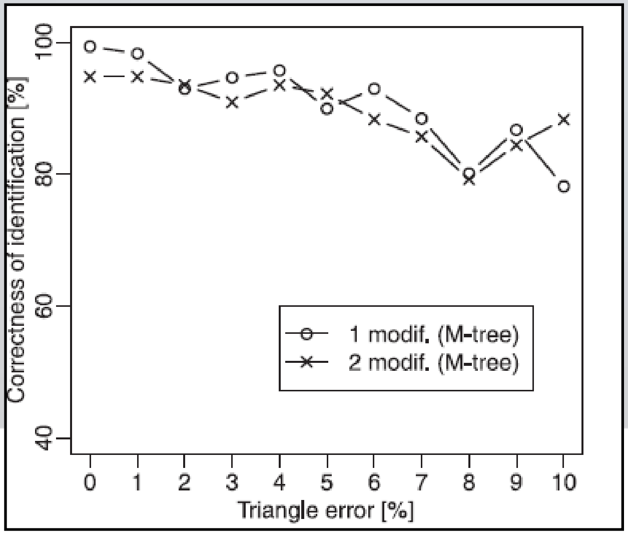


- d_{HP} – the indexability is better with increasing T-error tolerance
- d_A – about 35% of all pairwise distances in d_A=1 (uncorrectable)

Dealing with PTMs (in progress)







Conclusions

- parametrised Hausdorff distance (d_{HP})
  - models the similarity among spectra very well
  - can be utilized by MAMs

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

References

- J. Novák, D. Hoksza: Similarity Search and Posttranslational Modifications in Tandem Mass Spectra, accepted at IEEE BIBM 2010, Hong Kong, China

- J. Novák, T. Skopal, D. Hoksza, J. Lokoč: Improving the Similarity Search of Tandem Mass Spectra using Metric Access Methods, SISAP 2010, Istanbul, Turkey. ACM ISBN 978-1-4503-0420-7, pp. 85-92.

- J. Novák, D. Hoksza: Parametrised Hausdorff Distance as a Non-Metric Similarity Model for Tandem Mass Spectrometry, DATESO 2010, Štědronín - Plazy, Czech Republic. CEUR proceedings volume 567, ISSN 1613-0073, pp. 1-12.

- J. Novák, D. Hoksza: An Application of the Metric Access Methods to the Mass Spectrometry Data, IEEE CIBCB 2009, Nashville, Tennessee, USA.
  ISBN 978-1-4244-2756-7, pp. 220-227.