Protein Sequences Identification using NM-tree

Jiří Novák, Tomáš Skopal, David Hoksza, Jakub Lokoč and Jakub Galgonek
Charles University in Prague
Faculty of Mathematics and Physics
SIRET Research Group
http://siret.ms.mff.cuni.cz

ABSTRACT

We have generalized a method for tandem mass spectra interpretation, based on the parameterized Hausdorff distance d_{HP}. Instead of just peptides (short pieces of proteins), in this paper we describe the interpretation of whole protein sequences. For this purpose, we employ the recently introduced NM-tree to index the database of hypothetical mass spectra for exact or fast approximate search. The NM-tree combines the M-tree with the TriGen algorithm in a way that allows to dynamically control the retrieval precision at query time. A scheme for protein sequences identification using the NM-tree is proposed.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing—*indexing methods*

General Terms

Design, Performance

1. INTRODUCTION

Protein sequences are used in many fields of biological research, while tandem mass spectrometry is a fast and modern method for determining the sequences from an ”in vitro” sample. A mass spectrometer produces thousands of mass spectra for a few proteins in the sample. The proteins are split to many peptide ions, where a mass spectrum corresponds to a peptide ion. More peptide ions correspond to a peptide sequence. Multiple peptide sequences come from a protein sequence.

A mass spectrum (Fig. 1) is a list of peaks corresponding to peptide fragment ions. A peak is represented by the pair $\left(\frac{m}{z}, I\right)$, where $\frac{m}{z}$ is a mass-to-charge ratio and I is the intensity of a fragment ion occurrence.

The successful methods for mass spectra interpretation (i.e., matching the correct peptide sequences to the spectra) are based on the similarity search in databases of already known or theoretically predicted protein sequences. For this task, the non-metric parameterized Hausdorff distance d_{HP} (Eq. 2) and the technique for efficient search in a database of mass spectra indexed under d_{HP} were proposed [3], as

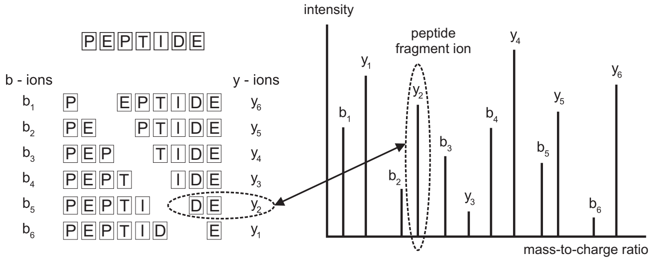

Figure 1: An example of a mass spectrum.

$$h(\vec{x},\vec{y}) = \frac{\sum_{\vec{x}_i \in \vec{x}} \sqrt[n]{\min_{\vec{y}_j \in \vec{y}} \{\max(0, |\vec{x}_i - \vec{y}_j| - \xi)\}}}{\dim(\vec{x})} \tag{1}$$

$$d_{HP}(\vec{x},\vec{y}) = \max(h(\vec{x},\vec{y}), h(\vec{y},\vec{x})) \tag{2}$$

where $\vec{x}$ and $\vec{y}$ represent vectors of real numbers ($\frac{m}{z}$ ratios), $\dim(\vec{x})$ is the length of $\vec{x}$ and ξ is a mass error tolerance. The technique employs the M-tree [1] for indexing hypothetical mass spectra generated from peptide sequences and prior to indexing the TriGen algorithm [4] is utilized to control the metricity of d_{HP}.

2. NM-TREE

The NM-tree (Non-Metric tree) [5] is a modification of the M-tree which natively aggregates the TriGen algorithm to support flexible approximate or exact search using an arbitrary (non)metric distance function. The approximate search is controlled by a modifier function f, e.g., the FP-modifier (Eq. 3), while the TriGen determines an optimal weight w for the specified T-error tolerance θ [4].

$$\mathrm{FP}(\delta, w) = \begin{cases} \delta^{\frac{1}{1+w}} & \text{for } w > 0 \\ \delta^{1-w} & \text{for } w \leq 0 \end{cases} \tag{3}$$

In the NM-tree, an input distance δ is supposed as a semi-metric, while TriGen is applied before indexing in order to turn δ into a metric δ^{f_M} (i.e., $\theta = 0$). Distances stored in the NM-tree are always the metric ones (i.e., $\delta^{f_M}(\cdot,\cdot)$). When a query (e.g., k-NN or range) is performed and the approximate search is required (i.e., $\theta > 0$), the distance δ^{f_M} is by definition modified inversely by f_M^{-1} and then another modifier f' is applied, i.e., $f'(f_M^{-1}(\delta^{f_M}))$ is computed. However,

SISAP '11, June 30-July 1, 2011, Lipari, Italy.
Copyright 2011 ACM 978-1-4503-0795-6/11/06 ...$10.00.

only distances at the pre-leaf and leaf level are modified by f_M^{-1} and f' (Fig. 2). The upper levels are not modified because the NM-tree stores not only direct distances between two objects (the to-parent distances) but also radii, which consist of aggregations [5].

A computation of modified distances can be expensive and it can degrade the overall NM-tree's performance. Nevertheless, this could be solved by a table of precomputed modified distances. Moreover, computation of ${d_{HP}}^{f_M}$ and $f_M^{-1}({d_{HP}}^{f_M})$ can be omitted for the purposes of mass spectra interpretation, as d_{HP} is already a metric distance.

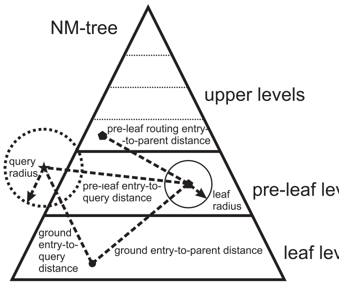

Figure 2: Distances modified in NM-tree.

3. APPLICATION

The above described NM-tree can be used with advantage in a scheme for protein sequences identification which is shown in Fig. 4. Let Q be a query set of peptide mass spectra coming from multiple protein sequences (e.g., 18 in our experiments). The goal is to identify protein sequences in Q using NM-tree with maximal θ minimizing the runtime. kNN queries (for random uninterpreted peptide spectra from Q) are performed using the NM-tree (Fig. 4a), while θ is being decreased (Fig. 4e) in order to improve the correctness. For each query, the k returned hypothetical spectra (peptide sequences) are matched to the original protein sequences (Fig. 4b). After the number of matched peptide sequences for a protein sequence exceeds a, the protein sequence is split to peptide sequences (Fig. 4c), i.e., all possible peptide sequences are generated. The generated peptide sequences are compared with all uninterpreted spectra in Q, the query spectrum becomes interpreted when a final ranking exceeds b (Fig. 4d).

The following experimental results consider single kNN queries (Fig. 4a), while the implementation of the rest of scheme is the subject of our future work. We compared the NM-tree with the set of M-trees for different T-error tolerances θ. We used kNN (k=1,000) queries, the database containing 100,000 protein sequences (5.6 million peptide seq.) and the query set containing 1,941 mass spectra [3]. The spectra in the query set come from 18 different proteins [2].

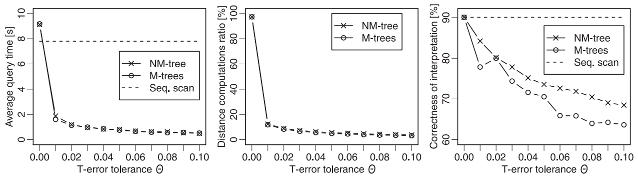

Figure 3: A comparison of the NM-tree with the set of M-trees for different T-error tolerances θ.

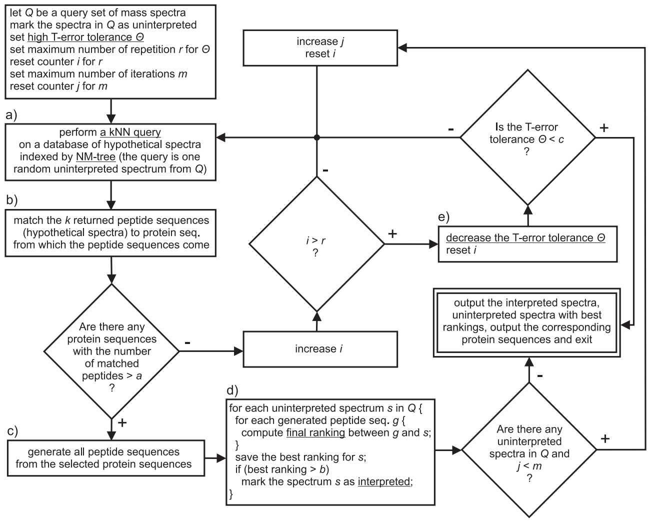
let Q be a query set of mass spectra
mark the spectra in Q as uninterpreted
set high T-error tolerance Θ
set maximum number of repetition r for Θ
reset counter i for r
set maximum number of iterations m
reset counter j for m
a)
perform a kNN query
on a database of hypothetical spectra
indexed by NM-tree (the query is one
random uninterpreted spectrum from Q)
b)
match the k returned peptide sequences
(hypothetical spectra) to protein seq.
from which the peptide sequences come
Are there any
protein sequences
with the number
of matched
peptides > a
?
-
+
c)
generate all peptide sequences
from the selected protein sequences
increase j
reset i
i > r
?
-
+
increase i
d)
for each uninterpreted spectrum s in Q {
for each generated peptide seq. g {
compute final ranking between g and s;
}
save the best ranking for s;
if (best ranking > b)
mark the spectrum s as interpreted;
}
Is the T-error
tolerance Θ < c
?
-
+
e)
decrease the T-error tolerance Θ
reset i
output the interpreted spectra,
uninterpreted spectra with best
rankings, output the corresponding
protein sequences and exit
Are there any
uninterpreted
spectra in Q and
j < m
?
-
+

Figure 4: Scheme for protein seq. identification.

The average query time and distance computations ratio (number of d_{HP} calls w.r.t. sequential scan) for the NM-tree are almost the same as for the set of M-trees (Fig. 3). The NM-tree is 15.6× faster than the sequential scan ($\theta = 0.1$). The search is faster with increasing θ, while the correctness of peptide sequences identification (or correctness of mass spectra interpretation) w.r.t. sequential scan is lower (Fig. 3c). However, low correctness of peptide sequences identification does not have to decrease the correctness of protein sequences identification because of more spectra for a peptide sequence and because of more peptide sequences in a protein sequence. The correctness of mass spectra interpretation is better for the NM-tree than for the set of M-trees with increasing θ.

Acknowledgments

This work has been supported by Czech Science Foundation (GAČR) projects 201/09/0683, 202/11/0968, 201/09/H057, by the Grant Agency of Charles University (GAUK) project Nr. 430711 and by the grant SVV-2011-263312.

4. REFERENCES

[1] P. Ciaccia, M. Patella, and P. Zezula. M-tree: An Efficient Access Method for Similarity Search in Metric Spaces. In *Proc. of 23rd Int. Conf. on VLDB*, pages 426–435, 1997.

[2] A. Keller and et al. Experimental Protein Mixture for Validating Tandem Mass Spectral Analysis. *OMICS: A Journal of Integrative Biology*, 6(2):207–212, 2002.

[3] J. Novák, T. Skopal, D. Hoksza, and J. Lokoč. Improving the Similarity Search of Tandem Mass Spectra using Metric Access Methods. In *SISAP '10*, pages 85–92, 2010.

[4] T. Skopal. Unified Framework for Fast Exact and Approximate Search in Dissimilarity Spaces. *ACM Transactions on Database Systems*, 32(4):29, 2007.

[5] T. Skopal and J. Lokoč. NM-Tree: Flexible Approximate Similarity Search in Metric and Non-metric Spaces. In *DEXA '08*, pages 312–325, 2008.