An Optimized Data Structure For High-Throughput 3D Proteomics Data: mzRTree

Nasso S, Silvestri F, Tisiot F, Di Camillo B, Pietracaprina A, Toffolo GM

Motivation

As an emerging field, mass spectrometry-based proteomics still requires software tools for efficiently storing and accessing experimental data. Here, we focus on the management of Liquid Chromatography-Mass Spectrometry (LC-MS) data, commonly available in standard XML-based formats. These formats can be highly computationally inefficient, especially when dealing with high-throughput profile data. LC-MS datasets are usually accessed through 2D range queries by means of either a m/z range, or a retention time range, or a combination of them, defining chromatograms, spectra, and peptide range queries, respectively. Optimizing them would dramatically improve the computational performance of data analysis. Our proposal is a scalable 2D indexing approach implemented through an R-tree-based data structure, called mzRTree, which can be efficiently built and stored and ensures efficient 1D and 2D data access.

Methods

mzRTree relies on a sparse matrix representation of the dataset, which is appropriate for MS-based proteomics data. In fact an MS 3D dataset (e.g. LC-MS), can be regarded as a 2D matrix and is characterized by a large number of null entries, whereas the non-null entries are distributed according to known patterns. In our approach, this matrix is divided into strips of consecutive rows and then each strip is partitioned into a suitable number of Bounding Boxes (BBs) containing non-zero entries. One file per strip is created, where all BBs belonging to the strip are stored. Therefore it can be efficiently loaded in the main memory during a range query. If less than a half entries of the BB have zero intensity, then the BB is stored using a dense matrix representation, instead of a sparse one. In order to efficiently support range query operations, we chose to implement the index through a balanced search tree based on the R-tree: leaves contain pointers to the BBs, while internal nodes are associated with larger bounding boxes including all BBs at the leaves of their respective sub-trees. We call mzRTree the whole data structure, which includes both the actual data (i.e., the BBs), and the tree index, and the metadata stored in a XML file compliant to mzXml/mzMl schema. A Java implementation of mzRTree is available at http://www.dei.unipd.it/mzrtree and

Department of Information Engineering, University of Padova, Padova, Italy

allows to build an mzRTree starting from an input dataset provided in mzXML/mzML format and to perform a generic range query.

Results

We evaluated the performances of mzRTree compared to Chrom (C) and Openraw (O), which are two existing intermediate formats used by Maspectras and MapQuant software respectively, both optimized for data access on one dimension: the former for chromatogram and the latter for spectra based access. As shown in the figure for a real profile 5 GB dataset, mzRTree outperforms C and O on all range queries and has fastest access times on the peptide query, which is likely to be the most common one in proteomic data analysis. Further results [Nasso et al., An optimized data structure for high-throughput 3D proteomics data: mzRTree, Journal of Proteomics, 2010] show that mzRTree requires the smallest hard disk space and data structure loading time, and it features an efficient creation time. Moreover, mzRTree is fairly scalable as regards access and data structure load time: as data density increases by a factor 10, the access time increases by a factor less than 3, while the load time is approximately constant. Experimental results and the R-tree structure scalability suggest that mzRTree is suitable for high density/large size proteomics data, such as profile data. Actually, profile data are often the only data source rich enough to carry on a meaningful analysis, e.g., in quantitative proteomics based on stable isotope labelling. However, computational costs involved with profile data handling often outweigh their benefits. mzRTree could revert this relationship.

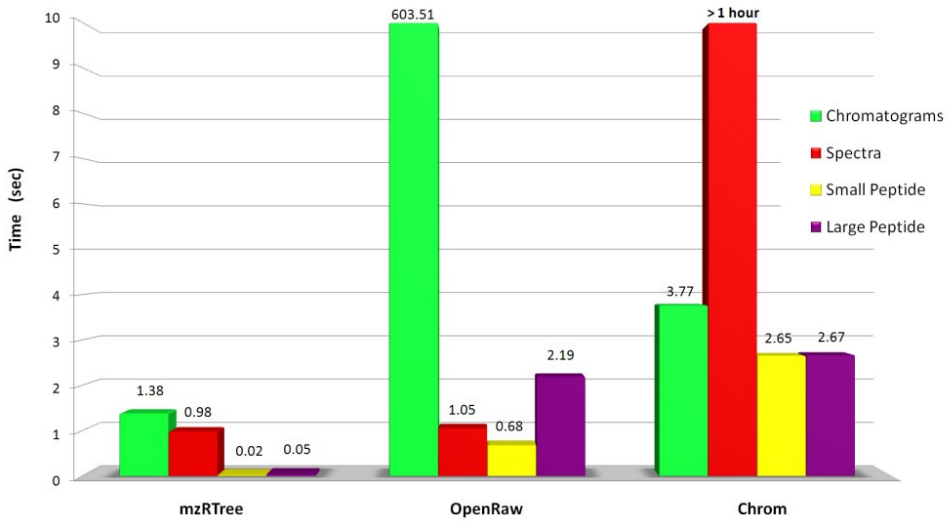
Availability

http://www.dei.unipd.it/mzRTree

Contact e-mail

mzrtree@dei.unipd.it

Image



Access times for a real profile 5GB dataset. Comparison among mzRTree, Chrom and OpenRaw on four different range queries: mzRTree reaches the best performance. Chromatograms refers to all retention times and a 5 Da range in the m/z dimension; spectra to the entire m/z dimension and 20 retention times; small and large peptide to 5 Da and 60 or 200 retention times, respectively. To avoid unintended locality effects, the access times required to perform ten range queries spanning the whole dataset are reported.