Title

A Web Service Framework for Interactive Analysis of Metabolomics Data

Permalink

https://escholarship.org/uc/item/7hw0n968

Journal

Analytical Chemistry, 89(11)

ISSN

0003-2700

Authors

Lyutvinskiy, Yaroslav
Watrous, Jeramie D
Jain, Mohit
et al.

Publication Date

2017-06-06

DOI

10.1021/acs.analchem.7b00890

Published in final edited form as:

Anal Chem. 2017 June 06; 89(11): 5713–5718. doi:10.1021/acs.analchem.7b00890.

A web service framework for interactive analysis of metabolomics data

Yaroslav Lyutvinskiy[1,2], **Jeramie Watrous**[3], **Mohit Jain**[3], and **Roland Nilsson**[1,2]

[1]Karolinska Institutet, Department of Medicine, Karolinska University Hospital, Stockholm, SE 17176

[2]Karolinska Institutet, Center for Molecular Medicine, Karolinska University Hospital, Stockholm, SE 17176

[3]UCSD, Chemistry and Biochemistry, 9500 Gilman Avenue, La Jolla, CA, USA, 92093

Abstract

Analyzing mass spectrometry-based metabolomics data presents a major challenge to metabolism researchers, as it requires downloading and processing large data volumes through complex “pipelines”, even in cases where only a single metabolite or peak is of interest. This presents a significant hurdle for data sharing, reanalysis or meta-analysis of existing data sets, whether locally stored or available from public repositories. Here we introduce mzAccess, a software system that provides interactive, online access to primary mass spectrometry data in real time via a web service protocol, circumventing the need for bulk data processing. mzAccess allows querying instrument data for spectra, chromatograms, or two-dimensional MZ-RT areas in either profile or centroid modes through a simple, uniform interface that is independent of vendor or instrument type. Using a cache mechanism, mzAccess achieves response times in the millisecond range for typical LC-MS peaks, enabling real-time browsing of large data sets with hundreds or even thousands of samples. By simplifying access to metabolite data, we hope that this system will help enable data sharing and reanalysis in the metabolomics field.

Introduction

Metabolomics is an increasingly important methodology in biological and biomedical sciences, given the renaissance in metabolism and biochemistry based research[1] and the expanding usage of clinical metabolomics data. In particular, liquid chromatography-mass spectrometry (LC-MS) based methods have developed rapidly over the past years, with greatly improved mass accuracy, sensitivity, and scan rates[2,3]. This creates exciting new opportunities for metabolomics research, but also leads to steadily growing data volumes. For example, a full-scan, high resolution LC-MS analysis with an Orbitrap mass

Correspondence to: Roland Nilsson.

Conflict of Interest Disclosure

Authors declare no competing financial interest.

Supporting Information

MzAccess web-service API reference is available free of charge via the Internet at http://pubs.acs.org/

spectrometer, a commonly used instrument, can easily produce 10Mb of data per minute of run time, so that 100 samples analyzed with a 10 minute LC-MS method will result in a 10Gb data set. At this rate, analyzing large sample collections can quickly result in data sets that are computationally and logistically difficult to handle.

A common paradigm for metabolomics data analysis is the "pipeline" model (Figure 1A), where primary instrument data is transformed through a series of steps into a condensed form (typically, peak areas) that can be further analyzed depending on the research question at hand. This approach is suitable for large-scale, untargeted data analysis involving thousands of peaks, and sophisticated software has been developed to handle tasks like peak picking, peak alignment, and normalization, including open source analysis suites such as XCMS, OpenMS and MZmine[4–8]. However, the pipeline approach is impractical for "interactive" analysis, where the user seeks to explore specific metabolites in a hypothesis-driven manner, since one must process gigabytes of data in order to extract a handful of LC-MS peaks of interest. Vendor software on the other hand is typically designed for interactive analysis, but is not widely available and offers only a limited set of functions. Moreover, reanalysis of published data is difficult and time-consuming with current data analysis software. Although metabolomics data repositories are now emerging[9,10], there is no facile way of accessing primary data without first downloading and re-processing an entire data set, even if only a single metabolite is of interest. This situation contrasts with that of sequence-based "omics" technologies, where genome-wide data is available online for thousands of datasets and can immediately be queried by gene symbols or sequence, greatly facilitating secondary analysis and data sharing[11,12]. An analogous mechanism for access to mass spectrometry-based metabolomics data would be valuable to promote data reanalysis, cross-comparisons and meta-analysis.

In this paper, we describe a new model for interactive, online analysis of metabolomics data (Figure 1B). Within this paradigm, direct, programmable access to primary mass spectrometry data is provided by an online service that users can browse in real-time for peak data of interest, thereby removing the need for data downloading and processing. As proof-of-principle, we present the mzAccess web service framework that implements the interactive model, and demonstrate its performance. With mzAccess, users can work with mass spectrometry data from any supported instrument or vendor in a uniform manner, from any internet-connected location and virtually any data analysis software. We anticipate that this model can be useful for public data repositories as well as local data servers in metabolomics labs or core facilities.

Methods

Software implementation

The mzAccess server software was implemented on a Microsoft Windows platform in C# 4.0 with Microsoft Visual Studio 2015, for use with the Microsoft Internet Information Service (IIS) web server software v7.5. Since IIS and the .NET framework contains its own parallelization mechanism, the web-service was designed to benefit from parallel requests processing. The server software reads primary data from instrument native formats (currently, Thermo .raw and Agilent .d formats supported) by interfacing with vendor APIs, and was designed in a modular fashion so that support for additional formats can be easily added as “plug-ins” by implementing a single C# class. Binaries and source code for the server software are freely available under the Apache License 2.0 for download from www.mzaccess.org.

The web service API was designed based on the SOAP (Simple Object Access Protocol) remote procedure call protocol, and consists of a small set of functions for reading chromatograms, spectra and MZ-RT areas. The web service was designed to be stateless (meaning that every service call is self-contained and does not require any previous or subsequent calls) in order to simplify simultaneous communication with multiple users.

The cache access mode was implemented by introducing an additional “cache” file for each primary data file, which contains centroid mode data only, sorted by MZ and indexed to fixed length data chunks. These cache files are generated automatically by an accompanying data conversion tool, which requires only an intensity threshold parameter that defines the minimum centroided peak intensity to include in the cache file. A full description of the web service API and the cache file structure is provided in Supporting materials.

Test environment and data

Performance test were carried out on a Dell PowerVault R610 server equipped with two Xeon E5645 CPUs (12 computational cores total), 32 Gb RAM and 4 WD20NPVX 2Tb HDD drives in software RAID 5 array, running Windows Server 2008 R2 64-bit operating system and IIS v7.5. The client computer was a laptop with Intel i7-2820QM CPU (4 cores) and 16 Gb RAM, running Windows 7 64-bit operating system. We tested the web service on two data sets, one acquired on a Thermo QExactive Orbitrap instrument (71 files) and one from an Agilent 6550 QTOF instrument (26 files).

For testing purposes, we generated a collection of MZ-RT coordinates of LC-MS chromatograms for these files using an in-house untargeted algorithm, which resulted in 20,870 chromatograms per file for the Thermo data set, and 99,576 for the Agilent data. We then generated client calls to the server by randomly selecting traces from this collection. Response time for each call was measured on the client side as the time interval from initiating a call until receiving complete data. We performed 100,000 test queries for chromatograms and 20,000 for MZ-RT areas (due to the longer response time in primary data mode in this case).

The IIS 7.5 web server software that hosts the mzAccess web service has its own parallelization mechanisms and can process multiple queries simultaneously in different threads. To investigate web service performance during intensive multi-user load, we emulated multiple users by separate instances of the client application running in parallel, each submitting 5000 service requests in sequence, and measured server CPU usage (using the operating system performance monitor) and response time as before. We tested up to 15 parallel instances, at which point the client computer CPU usage reached 70%, so that further testing was not possible. Due to this difficulty, the actual server load may have been slightly less than that of 15 fully independent clients.

Performance profiling was done using a single client submitting 20,000 requests for chromatograms in sequence. Two rounds of testing were performed, using either a 1Gbit Ethernet connection E (ping time ~1ms) or a 55Mbit WiFi connection W (ping time ~8ms), both for primary data and cache files. For both network conditions we estimated (1) network transfer time N_E and N_W, which consists of network latency time and data transfer time; (2) data retrieval time R_E and R_W, which is the time to read data from media and provide it to IIS for transfer to client side; (3) SOAP overhead time S_E and S_W, which comprises XML processing time on both client and server side plus overhead for IIS processing and for libraries and networking subsystem on both sides. Total response times for the mzAccess service A_E and A_W were measured as before, while data retrieval times R_E and R_W were measured directly on the server. In both cases, the data was binned with respect to chromatogram length and the median of each bin was taken as the final measure. Based on network speed tests, we also estimated that WiFi transfer time N_W is about 15 times the Ethernet time N_E. Finally, we assume that S_E and S_W are equal, since SOAP processing is independent of the network connection used. Hence, we have the equations

$$A_E=R_E+N_E+S_E$$

$$A_W=R_W+N_W+S_W$$

$$N_W=15N_E$$

$$S_W=S_E$$

By solving this system, we obtain $N_E = D_E / 14$, $N_W = 15\ D_W / 14$, $S_E = S_W = (15\ D_W - D_E) / 14$, where $D_E = A_E - R_E$ and $D_W = A_W - R_W$.

Results and discussion

Design and main features

To enable fast and easy access to large-scale mass spectrometry-based metabolomics data for end users, we chose a design where all data is hosted on a central server that users access over the internet via a web service (Figure 1B). This interactive model allows the bulk of sensitive primary data to be stored in a central and secure location, while end-users need only manage a small amount of processed data, such as peak areas for metabolites of interest. A typical metabolomics laboratory, core facility or public repository would maintain a central mzAccess server, while end users (lab members, collaborators, or external users of public data) only need a minimal amount of client software on their local computers to access the server.

To ensure reproducibility and prevent loss of information, mzAccess provides direct access to the primary, unprocessed instrument data via vendor-supplied software libraries. The system provides users with three common types of data: m/z spectra at a given retention time (RT), extracted ion current chromatograms at given a mass/charge ratio (MZ), and MZ-RT area data, which provides detailed information for deeper analysis of signals of interest (Figure 1C). Spectra and MZ-RT areas can be accessed in both centroid and profile mode. It is also possible to extract MS^2, MS^3, or higher order fragmentation events, provided that the underlying file format supports this. In the current version, Thermo and Agilent formats are supported by the mzAccess server software, while support for other data formats can be added as needed in the form of plug-ins. Since all major instrument vendors only support the Windows platform, mzAccess servers must be Windows machines. However, client software can easily be implemented on any computer platform and in any software environment. At present, clients for the R programming language and the Mathematica computing platform are available from www.mzaccess.org.

As an example of typical interactive usage of the mzAccess system using the R client, Figure 2 shows an analysis of glutamate in extracts from cultured cells labeled with a $U^{13}C,^{15}N$glutamine tracer, analyzed on a Thermo QExactive Orbitrap instrument. To identify the retention time of glutamate in this data set, we first retrieved a full-length chromatogram at the $[\text{glutamate} + H]^+$ ion m/z from a standard mixture sample containing glutamate, revealing a single peak at 8.3 minutes (Figure 2A). To confirm the identity of this peak, we next viewed chromatograms from cell extracts at this retention time window, and observed a clear $^{13}C_5$ mass isotopomer peak in extracts from labeled cells, indicating that this species is indeed glutamate (Figure 2B). To verify separation of carbon and nitrogen mass isotopomers (which differ by 0.006u), we retrieved profile mode m/z spectra at the peak retention time, and observed two clearly resolved m/z peaks (Figure 2C). This is also shown across the elution of the chromatographic peak using MZ-RT area data in profile mode (Figure 2D).

mzAccess also provides access to MS/MS spectra in any MZ-RT area of interest. Figure 2E shows an example of analysis of bilirubin in a human plasma sample analyzed on an Agilent QTOF instrument. Here, we first checked the bilirubin MS^1 chromatogram for presence of MS/MS fragmentation events, and then retrieved one MS/MS spectra of interest, which indeed exhibits peaks indicating fragments of bilirubin. These analyses are instantly accessible from the web service, in this case using a few lines of R code (Figure 2), and require no preparation by the user other than knowing the MZ-RT coordinates of interest. R client including code mentioned on Figure 2 is available and test server to access the data is up and running (see mzAccess.org for links to resources).

Server performance

For interactive analysis of metabolomics data, it is crucial to achieve fast server response times, even for very large datasets. While vendor data formats offer the most accurate data representation and provide fast access to individual spectra, they are typically not optimized for fast access to chromatograms and MZ-RT area data[13]. We therefore developed an indexing/caching mechanism, here referred to as “cache mode”, which is used for most interactive browsing, while primary data is used only when maximal accuracy is needed. In cache mode, accessing a typical 1 minute chromatogram (~100 scans) takes about 10 ms, increasing linearly with chromatogram length (Figure 3A). The effect of caching becomes noticeable for long chromatograms, with cache mode being ~5 times faster than primary data access for a 25 minute chromatogram. For MZ-RT area queries, cache mode is essential, being ~100 times faster than primary data access (Figure 3B). Moreover, cache mode is optimized for accessing the same feature across a large number of files, which can give substantial performance gains. For example, accessing a single metabolite peak, or even multiple isotopomers of a metabolite, across more than 400 files is accomplished on the order of 1 second (Table 1). These results demonstrate that mzAccess allows browsing of metabolomics data in real time, even in large data sets.

For large laboratories and public data repositories, it is vital that servers maintain fast response times also when multiple users are simultaneously accessing data. In benchmarking experiments with up to 15 client programs continually requesting data from the server, mzAccess consistently maintained less than 10 ms response time in cache mode (Figure 4). This is because the caching mechanism of cache mode makes use of multiple processor cores: the server used for testing had 12 cores, and reached only 20% processor load at 15 simultaneous clients (Figure 4). These results indicate that the mzAccess system can support large multi-user environments.

Finally, we determined the network transfer speed required for optimal performance of mzAccess. We performed in-depth performance profiling of chromatogram access in cache mode on 1Gbps and 55Mbps connections, measuring the contribution of mzAccess processing time and network transfer to the total response time (Figure 5). On the 1Gbit network, processing time was the major component; however, this was in large part due to overhead of the SOAP web service protocol, suggesting that further optimization of our server software would not substantially improve overall performance. On the 55Mbps connection, transfer time was comparable to mzAccess processing time, indicating that this network speed is sufficient for optimal performance. Hence, mzAccess can be used with most modern internet connections, and can therefore provide immediate access to metabolomics data from any location.

Conclusion

Herein we have introduced the mzAccess web service framework and demonstrated that it achieves the performance necessary to support interactive online analysis of metabolomics data. We believe this approach fills a need for fast, easy access to mass spectrometry data for non-expert users, particular when a small number of metabolites are of interest. This approach is complementary to, but does not replace, the “pipeline” approach (Figure 1), which is more suitable when very large amounts of data must processed. We envision that the interactive approach could have an important role in enabling direct access to published data sets for data sharing, reanalysis and meta-analysis purposes, which is less developed in the metabolomics community today.

The mzAccess system is open software, publicly available at www.mzaccess.org, and is designed to be extensible to new data formats and new computational environments. The central web service protocol (see Supplementary material) specifies only the fundamental operations for requesting spectra, chromatograms and MZ-RT areas, and we have strived to avoid any unnecessary assumptions that might not hold for future instruments. The simplicity of the protocol makes it easy to develop client software for various platforms, ranging from programming languages to graphical user interfaces. On the server side, support for any instrument data format can be added in the form of plug-ins (Windows DLL files), which can be developed independently by advanced users. While we prefer using primary vendor-format instrument data whenever possible for the sake of accuracy, we will provide plug-ins for the commonly used mzML and mzXML formats to handle cases where the original data is no longer accessible.

Finally, we anticipate that mzAccess might serve as a foundation on which to build more sophisticated, higher-level data services. It must be emphasized that our web service only provides the most fundamental level of access to mass spectrometry data files, assuming that the user knows MZ-RT coordinates for metabolites of interest. However, it is easy to imagine combining mzAccess with established mass spectrometry databases[10,14] that can automatically provide such coordinates for a given metabolite on a specific instrument and analytical method, enabling analysis of metabolomics datasets by non-expert users. Similarly, our framework could be integrated with sample annotation databases[9,10] to automatically retrieve data from the relevant files according to the experimental design of a dataset. We hope that our contribution will help stimulate development of such services, and promote data exchange, exploration and re-use in the metabolomics community.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by grants from the Doris Duke Charitable Foundation (#2015092: M.J.), Tobacco-Related Disease Research Program (#24RT-0032: M.J., #24FT-0010: J.D.W.), and National Institutes of Health (1R01ES027595: M.J.), and the Swedish Foundation for Strategic Research (FFL12-0220.006; Y.L, R.N).

References

1. McKnight SL. Science. 2010; 330(6009):1338–1339. [PubMed: 21127243]
2. Junot C, Fenaille F, Colsch B, Bécher F. Mass Spectrom. Rev. 2014; 33(6):471–500. [PubMed: 24288070]
3. May JC, McLean JA. Annu. Rev. Anal. Chem. 2016; 9(1):387–409.
4. Pluskal T, Castillo S, Villar-briones A, Oresic M. BMC Bioinformatics. 2010; 11:395. [PubMed: 20650010]
5. Smith C, Want EJ, O’Maille G, Abagyan R, Siuzdak G. Anal. Chem. 2006; 78(3):779–787. [PubMed: 16448051]
6. Huang X, Chen Y-J, Cho K, Nikolskiy I, Crawford P, Patti GJ. Anal. Chem. 2014; 86(3):1632–1639. [PubMed: 24397582]
7. Katajamaa M, Oresic M. J. Chromatogr. A. 2007; 1158(1–2):318–328. [PubMed: 17466315]
8. Rost HL, Sachsenberg T, Aiche S, Bielow C, Weisser H, Aicheler F, Andreotti S, Ehrlich HC, Gutenbrunner P, Kenar E, et al. Nat. Methods. 2016; 13(9):741–748. [PubMed: 27575624]

9. Haug K, Salek RM, Conesa P, Hastings J, de Matos P, Rijnbeek M, Mahendraker T, Williams M, Neumann S, Rocca-Serra P, et al. Nucleic Acids Res. 2013; 41:D781–D786. [PubMed: 23109552]

10. Sud M, Fahy E, Cotter D, Azam K, Vadivelu I, Burant C, Edison A, Fiehn O, Higashi R, Nair KS, et al. Nucleic Acids Res. 2016; 44:D463–D470. [PubMed: 26467476]

11. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Holko M, et al. Nucleic Acids Res. 2013; 41:D991–D995. [PubMed: 23193258]

12. Kolesnikov N, Hastings E, Keays M, Melnichuk O, Tang YA, Williams E, Dylag M, Kurbatova N, Brandizi M, Burdett T, et al. Nucleic Acids Res. 2015; 43:D1113–D1116. [PubMed: 25361974]

13. Bouyssié D, Dubois M, Nasso S, Gonzalez de Peredo A, Burlet-Schiltz O, Aebersold R, Monsarrat B. Mol. Cell. Proteomics. 2015; 14(3):771–781. [PubMed: 25505153]

14. Smith CA, O'Maille G, Want EJ, Qin C, Trauger SA, Brandon TR, Custodio DE, Abagyan R, Siuzdak G. Ther. Drug Monit. 2005; 27(6):747–751. [PubMed: 16404815]

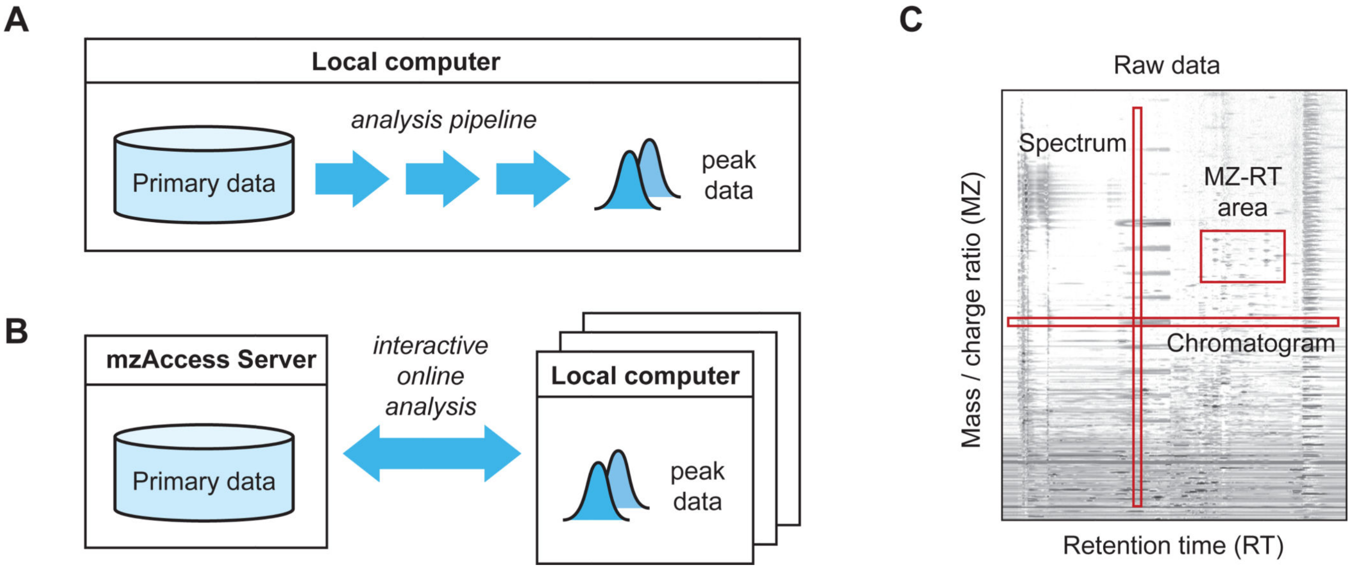
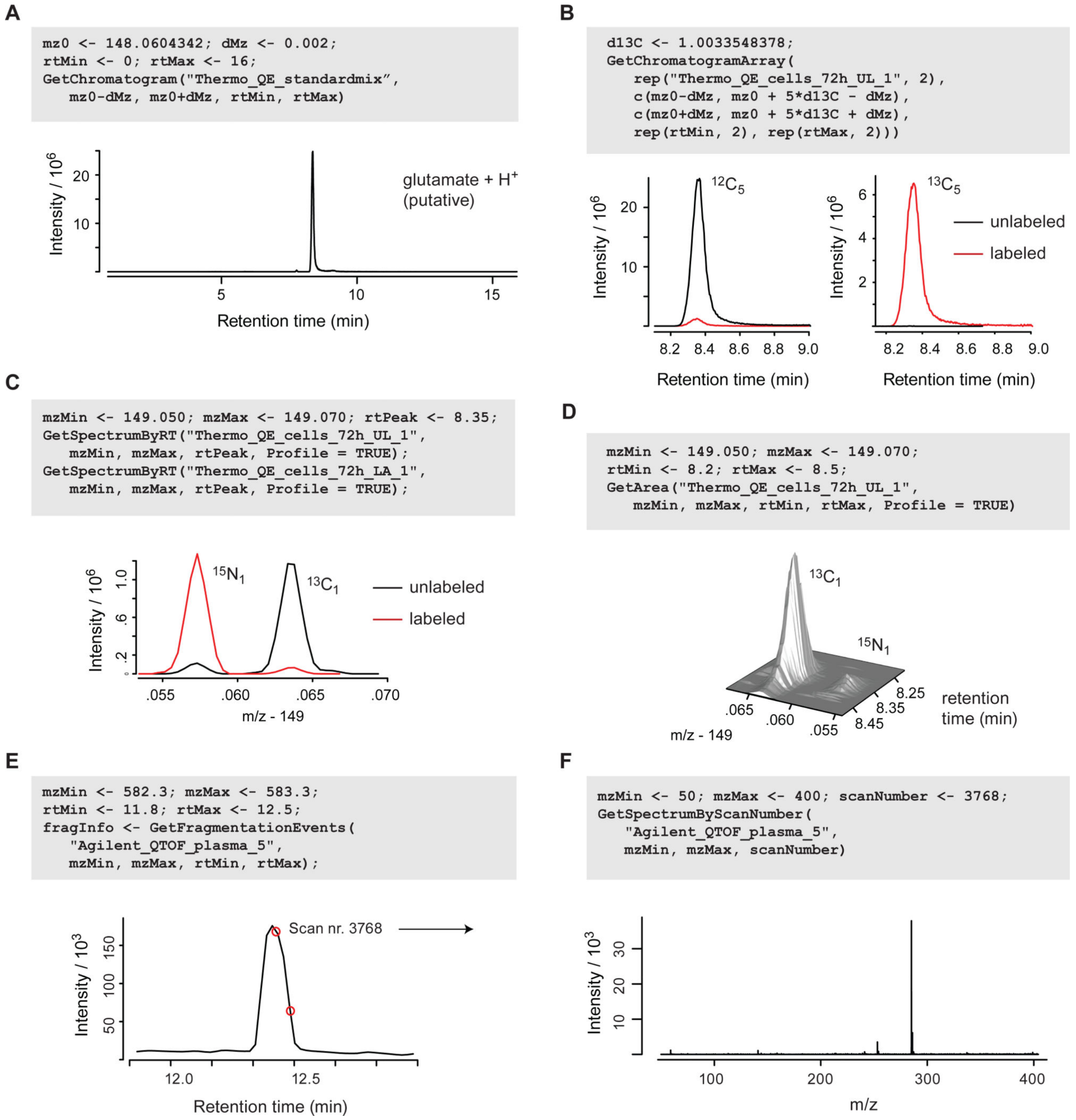


Figure 1. The interactive metabolomics data analysis

(A) In the pipeline model for metabolomics data processing, primary data is stored on a local computer and processed in a programmatic fashion in several steps, according to various parameters, to obtain peak data. (B) In the interactive model, primary data is hosted on a server, which provides peak data to clients on demand. (C) The main types of data provided by the mzAccess framework, illustrated on a full scan LC-MS data set.



```
mz0 <- 148.0604342; dMz <- 0.002;
rtMin <- 0; rtMax <- 16;
GetChromatogram("Thermo_QE_standardmix",
   mz0-dMz, mz0+dMz, rtMin, rtMax)
```

```
d13C <- 1.0033548378;
GetChromatogramArray(
   rep("Thermo_QE_cells_72h_UL_1", 2),
   c(mz0-dMz, mz0 + 5*d13C - dMz),
   c(mz0+dMz, mz0 + 5*d13C + dMz),
   rep(rtMin, 2), rep(rtMax, 2)))
```

```
mzMin <- 149.050; mzMax <- 149.070; rtPeak <- 8.35;
GetSpectrumByRT("Thermo_QE_cells_72h_UL_1",
   mzMin, mzMax, rtPeak, Profile = TRUE);
GetSpectrumByRT("Thermo_QE_cells_72h_LA_1",
   mzMin, mzMax, rtPeak, Profile = TRUE);
```

```
mzMin <- 149.050; mzMax <- 149.070;
rtMin <- 8.2; rtMax <- 8.5;
GetArea("Thermo_QE_cells_72h_UL_1",
   mzMin, mzMax, rtMin, rtMax, Profile = TRUE)
```

```
mzMin <- 582.3; mzMax <- 583.3;
rtMin <- 11.8; rtMax <- 12.5;
fragInfo <- GetFragmentationEvents(
   "Agilent_QTOF_plasma_5",
   mzMin, mzMax, rtMin, rtMax);
```

```
mzMin <- 50; mzMax <- 400; scanNumber <- 3768;
GetSpectrumByScanNumber(
   "Agilent_QTOF_plasma_5",
   mzMin, mzMax, scanNumber)
```

Figure 2. Example use case

(A) Chromatogram at the [glutamate + H$^+$] ion m/z, extracted from an analysis of a standard mixture on an LC-MS instrument with a 16 min chromatography method. (B) Chromatograms for the putative [glutamate + H$^+$] ion from analyses of extracts of cells cultured in unlabeled (black lines) and U-^{13}C-glucose, U-^{15}C,^{15}N-glutamine labeled medium (red lines). Left, chromatograms at the $^{12}C_5$ (base) isotopomer m/z; right, chromatograms at the $^{13}C_5$ (fully carbon-labeled) isotopomer m/z. Peak heights were normalized to total peak area over all isotopomers. (C) Single-scan spectra for same samples as in B, at RT = 8.35 minutes, for the indicated m/z range. Present mass isotopomers are indicated. (D) RT-MZ area accessed in profile mode for one unlabeled cell extract, with mass isotopomers indicated. (E) MS/MS fragmentation events (red circles) overlaid on a chromatogram of m/z =583.25Da. (F) Selected MS/MS spectrum for precursor mass of 583.25Da.

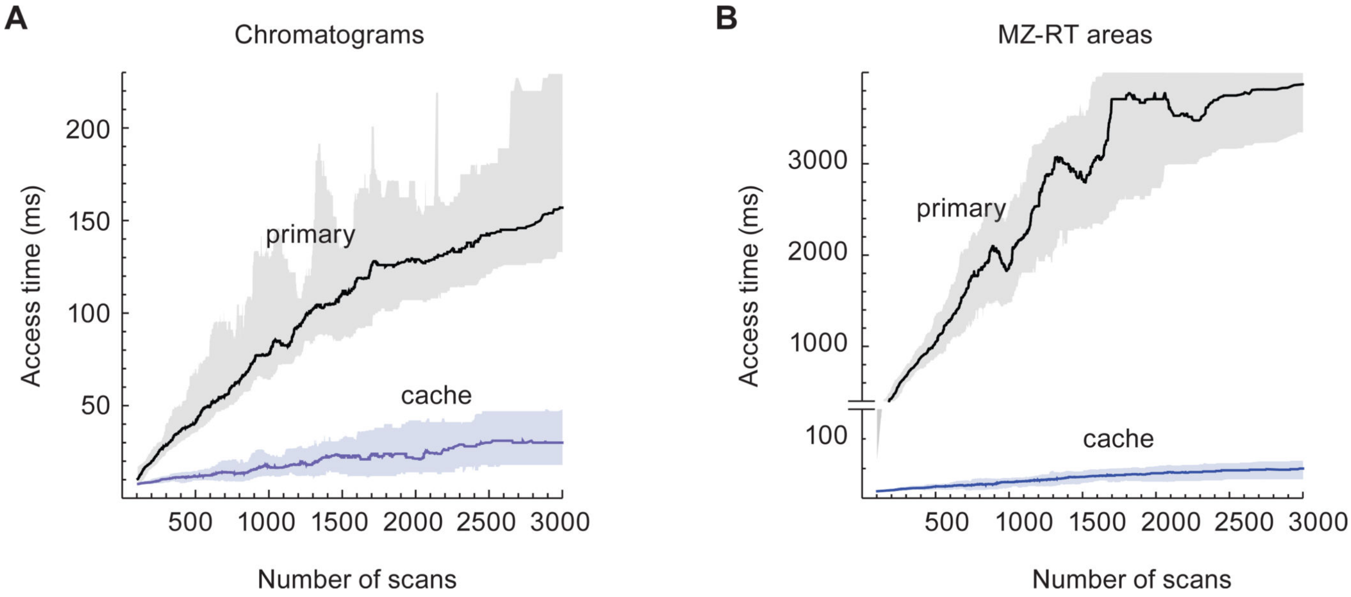


Figure 3. Performance of the mzAccess server

(A) Measured access times for chromatograms of indicated length using the primary (black) and cache (blue) access modes. Solid lines indicate running medians, while shaded area indicate running 10% and 90% percentiles, with a window size of 100 scans. (B) Measured access times for MZ-RT areas, as in A.

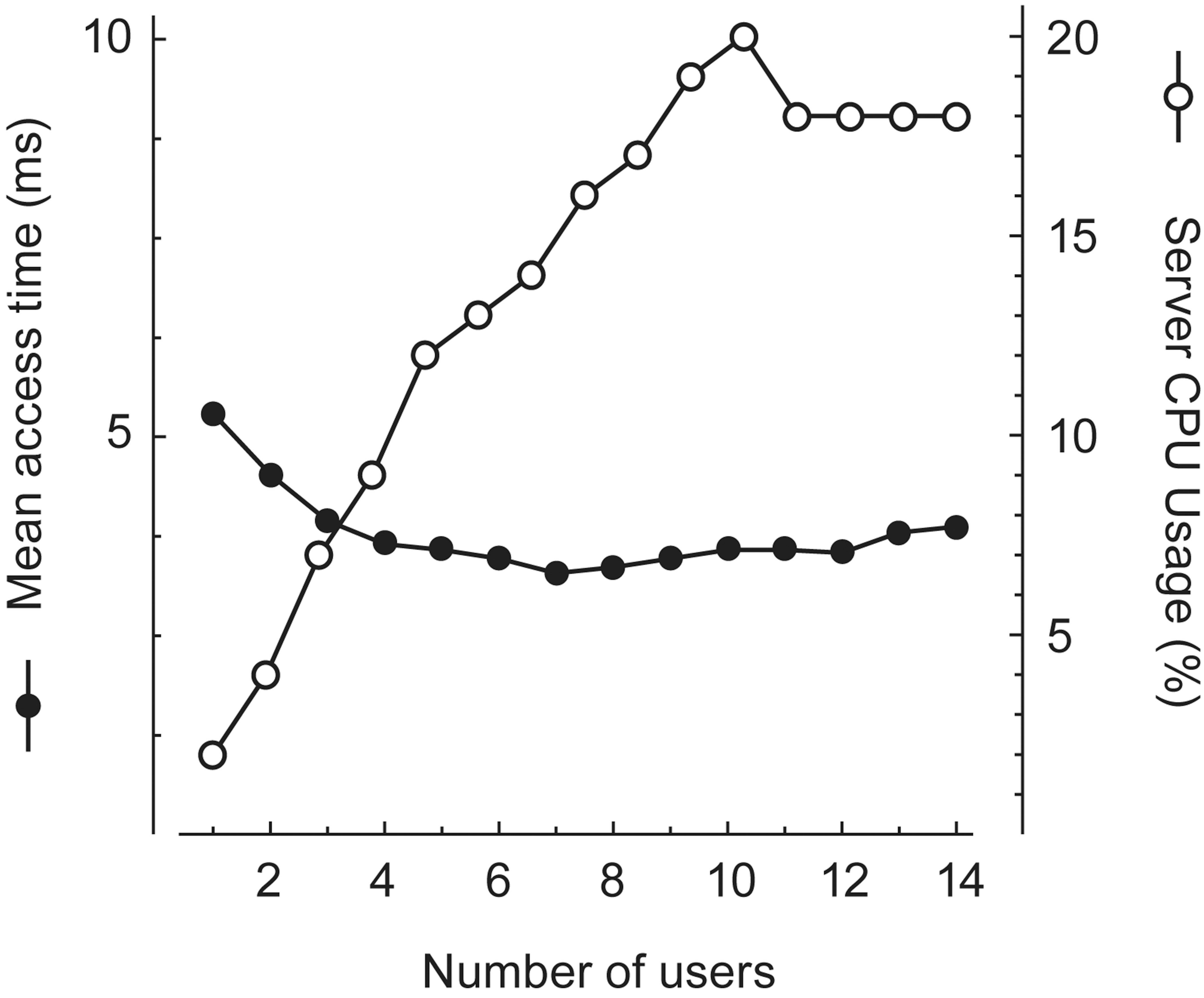


Figure 4. Server performance in multi-user context

Access times (filled circles) and server CPU usage (open circles) in cache mode, averaged over all data requests, for a varying number of users (here simulated as client processes).

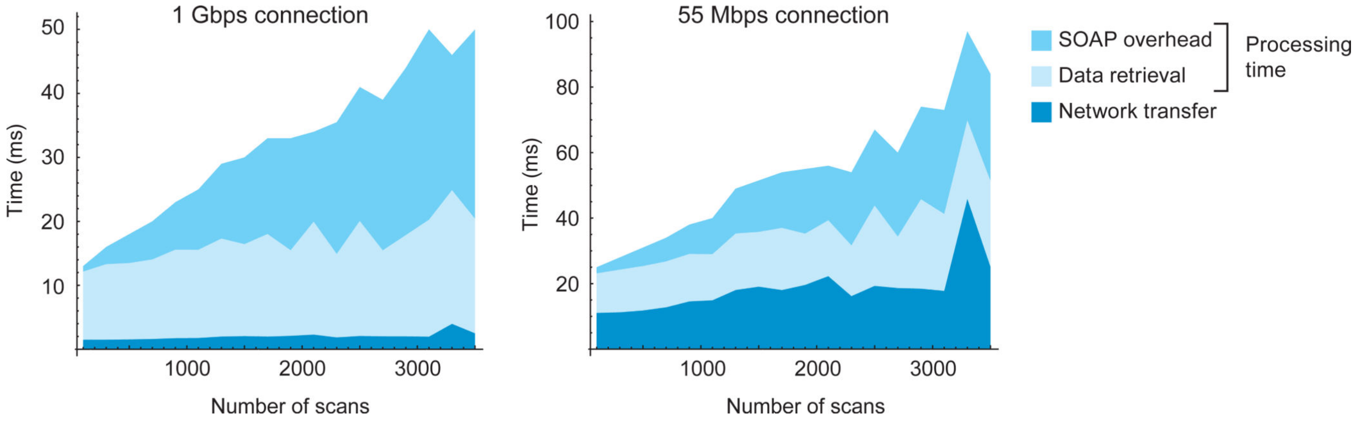


Figure 5. Profiling of server performance

Median access times in cache mode, partitioned into SOAP overhead, data retrieval time, and network transfer time, for chromatograms of indicated length, using a 1 Gbps network connection (left), or a 55 Mbps network connection (right).

Table 1

Access times for large data sets

A set of 420 and 439 data files collected in negative and positive ion modes, respectively, was queried for the metabolite ions listed. The total access time varies with the amount of data present. A data point is one (mz, intensity) pair in chromatogram data.

Metabolite	Adduct	No. mass isotopomers	Total no. chromatograms	RT interval (min)	Total access time (sec)	Total no. data points transferred
Glucose	H−	1	420	9.31–10.07	3.209	44744
Glucose	H−	5	2100	9.31–10.07	3.407	80912
Glutamate	H+	1	439	11.09–11.97	0.612	46918
Glutamate	H+	5	2195	11.09–11.97	2.865	104940
AMP	H+	1	439	12.27–12.74	0.256	13422
AMP	H+	5	2195	12.27–12.74	0.481	17978