

Software

Open Access

BioSunMS: a plug-in-based software for the management of patients information and the analysis of peptide profiles from mass spectrometry

Yuan Cao^{†1}, Na Wang^{†2}, Xiaomin Ying¹, Ailing Li², Hengsha Wang², Xuemin Zhang^{*2} and Wuju Li^{*1}

Address: ¹Center of Computational Biology, Beijing Institute of Basic Medical Sciences, Taiping Road 27, Haidian district, Beijing 100850, PR China and ²National Center of Biomedical Analysis, Taiping Road 27, Haidian district, Beijing 100850, PR China

Email: Yuan Cao - labs.net@gmail.com; Na Wang - wn@proteomics.cn; Xiaomin Ying - yingxm@nic.bmi.ac.cn; Ailing Li - lal@proteomics.cn; Hengsha Wang - wanghengsha@hotmail.com; Xuemin Zhang* - xmzhang@nic.bmi.ac.cn; Wuju Li* - liwj@nic.bmi.ac.cn

* Corresponding authors †Equal contributors

Published: 17 February 2009

Received: 2 September 2008

BMC Medical Informatics and Decision Making 2009, 9:13 doi:10.1186/1472-6947-9-13

Accepted: 17 February 2009

This article is available from: <http://www.biomedcentral.com/1472-6947/9/13>

© 2009 Cao et al; licensee BioMed Central Ltd.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/2.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract

Background: With wide applications of matrix-assisted laser desorption/ionization time-of-flight mass spectrometry (MALDI-TOF MS) and surface-enhanced laser desorption/ionization time-of-flight mass spectrometry (SELDI-TOF MS), statistical comparison of serum peptide profiles and management of patients information play an important role in clinical studies, such as early diagnosis, personalized medicine and biomarker discovery. However, current available software tools mainly focused on data analysis rather than providing a flexible platform for both the management of patients information and mass spectrometry (MS) data analysis.

Results: Here we presented a plug-in-based software, BioSunMS, for both the management of patients information and serum peptide profiles-based statistical analysis. By integrating all functions into a user-friendly desktop application, BioSunMS provided a comprehensive solution for clinical researchers without any knowledge in programming, as well as a plug-in architecture platform with the possibility for developers to add or modify functions without need to recompile the entire application.

Conclusion: BioSunMS provides a plug-in-based solution for managing, analyzing, and sharing high volumes of MALDI-TOF or SELDI-TOF MS data. The software is freely distributed under GNU General Public License (GPL) and can be downloaded from <http://sourceforge.net/projects/biosunms/>.

Background

With wide applications of MALDI-TOF MS and SELDI-TOF MS in biomedical studies, more and more large-scale MS datasets have being obtained [1-6]. How to extract useful information from these datasets not only needs a variety of statistical analysis, but also asks for patients

information. Thus, an efficient and flexible software is needed to comprehensively handle so much information and so many analytical tools. Up to now, some software has been developed for MS data analysis. For example, the database built by Titulaer [7] can analyze high-throughput MS data from MALDI-TOF MS measurements. But it

can not manage and analyze the clinical information related to the MS experimental results. In addition, Josep et al have developed a laboratory information management system for characterizing and standardizing the entire sample collection and serum preparation process [8]. The data processing and analysis were done in MATLAB [9] and GeneSpring [10]. Recently, Marc Sturm group presented a software framework, OpenMS, for rapid application development in mass spectrometry [11]. However, it is not specific for the analysis of MALDI-TOF or SELDI-TOF MS. Furthermore, there are many other projects for MS data processing and analysis e.g. MapQuant [12], MASPECTRAS [13], SpecArray [14], msInspect [15] and MZMine [16]. But few projects try to emphasize both the management of patients information and MALDI-TOF or SELDI-TOF MS-related statistical analysis. Moreover, none of them provides an integrated solution for clinical researchers without any knowledge in programming, as well as a plug-in architecture platform with the possibility for developers to add or modify functions without need to recompile the entire application.

Here we developed a flexible and compact software, BioSunMS, for MALDI-TOF or SELDI-TOF MS-based clinical proteomics study. The name BioSunMS was coined by the combination of BioSun and MS (mass spectroemtry), in which BioSun stands for the comprehensive bioinformatics software developed by our center [17]. BioSunMS was designed to support decision-making and allow patients information and spectra data to be stored, managed, processed, and analyzed using the Rich Client Platform (RCP) [18] from Eclipse [19]. The BioSunMS software had been tested with MS files of serum samples from patients with lung cancer and a control group.

Implementation

System architecture

BioSunMS provides a relational database and client-server architecture suitable for multi-user workgroups. It is a RCP application extending the Eclipse framework. It was developed in Java, and provided a domain-specific platform where the plug-ins can be integrated. End-users can select related features from the BioSunMS plug-ins freely.

BioSunMS system includes four functional modules, namely, data management, spectrum processing, MS profile analysis and security module (Figure 1). The data management module provides a robust, client-server, relational database system for the management of patients information and MS data. Data are primarily stored in a relational database MySQL, while some raw data is stored on the file system (mainly to provide backup for MS data). The spectrum processing module performs spectrum import, spectrum export, and related analysis such as calibration, normalization and peak

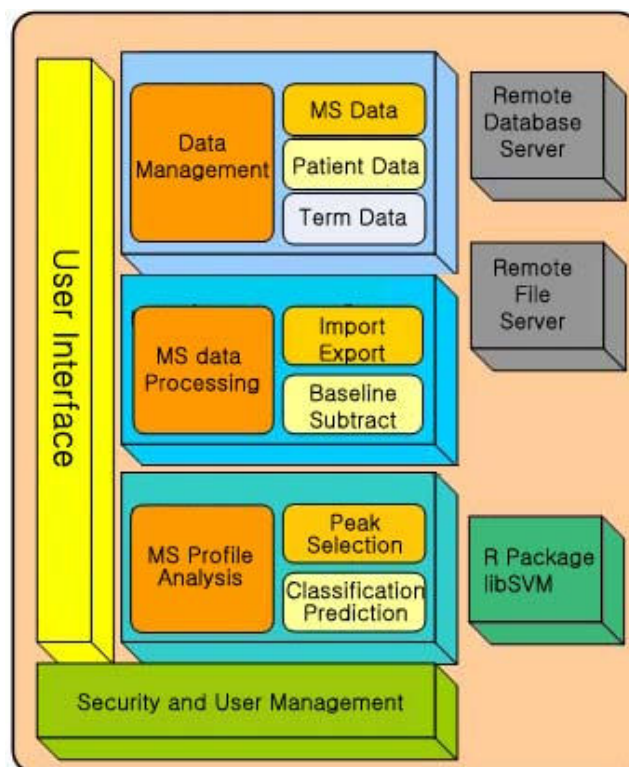


Figure 1

The main components of the BioSunMS. Four components are provided in BioSunMS system, which are Security and User Management, Data Management, MS data Processing, and MS profile Analysis, respectively.

detection [20]. The MS profile module is designed for sample classification and identification of potential biomarkers. Background computations are mostly done by R package and libSVM [21]. BioSunMS has a security system for protecting the data. Access to any data can be controlled. Access is granted via user groups rather than individual users. The main features of BioSunMS can be found in Table 1.

User Interface

BioSunMS is built using Standard Widget Toolkit (SWT). Created as part of the Eclipse project, SWT allows developers to build efficient and portable applications. In contrast to Swing/AWT (which provide their own graphical environment), SWT has the same look and feel of the operating system on which the application runs. AWT/Swing components can be wrapped in SWT, and this feature is utilized in BioSunMS to integrate Java components built on these toolkits [22]. In an Eclipse RCP application, the user interface is composed of five main graphical units, namely, View, Editor, Perspective, Menu and Wizard. A View is a graphical window to provide information for the users. An Editor is another type of window to let the users

Table 1: The main features of BioSunMS software.

Feature Item	Feature description
Database	MySQL based relational database system for patients information and MS data.
Preprocessing	Baseline subtraction and smoothing; Normalization; Automatic or manual peak detection; Peak alignment
MS format	Support common MS format, such as mzML, mzXML, mzData,
Visualization	Graphics display for MS raw data and MS peaks
Statistics or procedures	t-test for feature selection; SVM or kNN-based model construction;
ROC	The ROC curve for the evaluation of model performance
Security	Client-server architecture for multi-user workgroups
Scalability	Plug-in framework for future development

edit and save data. A perspective is a visual container for a set of views and content editors. An example in BioSunMS is the Spectral Perspective, which displays Views, Editors, and Menu options for spectral data. Wizards are used to guide users to finish particular object through a sequential set of tasks. The internal placement and size of components within a perspective are not fixed but can be changed at the user preference and is saved between sessions.

Workflow

To access the BioSunMS server at client side, users must first log on with an authorized username and a given password. Users could also log on for analyzing spectral data without connecting to the server, when their computers are offline. The workflow of BioSunMS was illustrated in Figure 2, which was divided into four main parts as follows.

Enter new records

The first step in the workflow is to input patients general information and laboratory test results. To add a new record, click on the top-right button "Append" in the toolbar. A form will be displayed to enter data according to the module. Fields include equipment, sample ID, gender, age, specimen group and so on. Marked fields must be specified. Users could group samples with the Query View.

Import spectra to the BioSunMS

The second step in the workflow is to import spectral files to the BioSunMS. After importing a spectral file, BioSunMS uses the Bioconductor PROcess package developed by Li [23] to process the spectral data. The PROcess package contains a collection of functions for removing baseline drifts, detecting peaks and aligning peaks for a set of biomarkers. In the Spectra View which uses the JFree-Chart package [24] for visualization of spectral information, users can identify or erase the peaks in a spectrum manually. Labeled peaks are shown in the Peak View. After data processing, the peak list can be saved to the database corresponding to the sample ID (Figure 3). Users

can upload the "raw" spectral files to the ftp server for backup.

Group spectra into folders

The third step in the workflow is to group spectral files into the Bioresource folder. To analyze data, spectra must be grouped together into a folder. There are many ways to organize spectra by querying the database for spectra meeting desired criteria, such as research group, user, sample state, sample type, patient and characteristic description. Users can also define the condition by selecting a series of fields to be used to separate spectra into groups for classification and prediction. After grouping, a web page with a table of characteristics of patients can be created (Figure 4).

Apply analysis algorithms

The fourth step in the workflow is to analyze the experimental results for identification of potential biomarkers and early diagnosis. BioSunMS provides the Wizard for general users, including extraction of MS peaks, construction of m/z matrix, sample classification and sample prediction. Advanced users can process and analyze the MS data in the Data Analysis Perspective. After building a Support Vector Machine (SVM) model on the training dataset and make prediction on the test dataset, BioSunMS uses a Receiver Operation Characteristic (ROC) curve to assess the models performance (Figure 5).

Results and discussion

BioSunMS is a RCP application through bringing the power of lab's network to access and manage a large amount of information. It can be installed on Windows or Linux system. It allows researchers to enter patient data in a customizable template and group the data by queries. It has a security system for protecting data. Access is granted via user groups rather than individual users.

Security and access

BioSunMS is a network-based system like an Intranet. It can be accessed from all computers from the same local network. It can even be accessed through the internet.

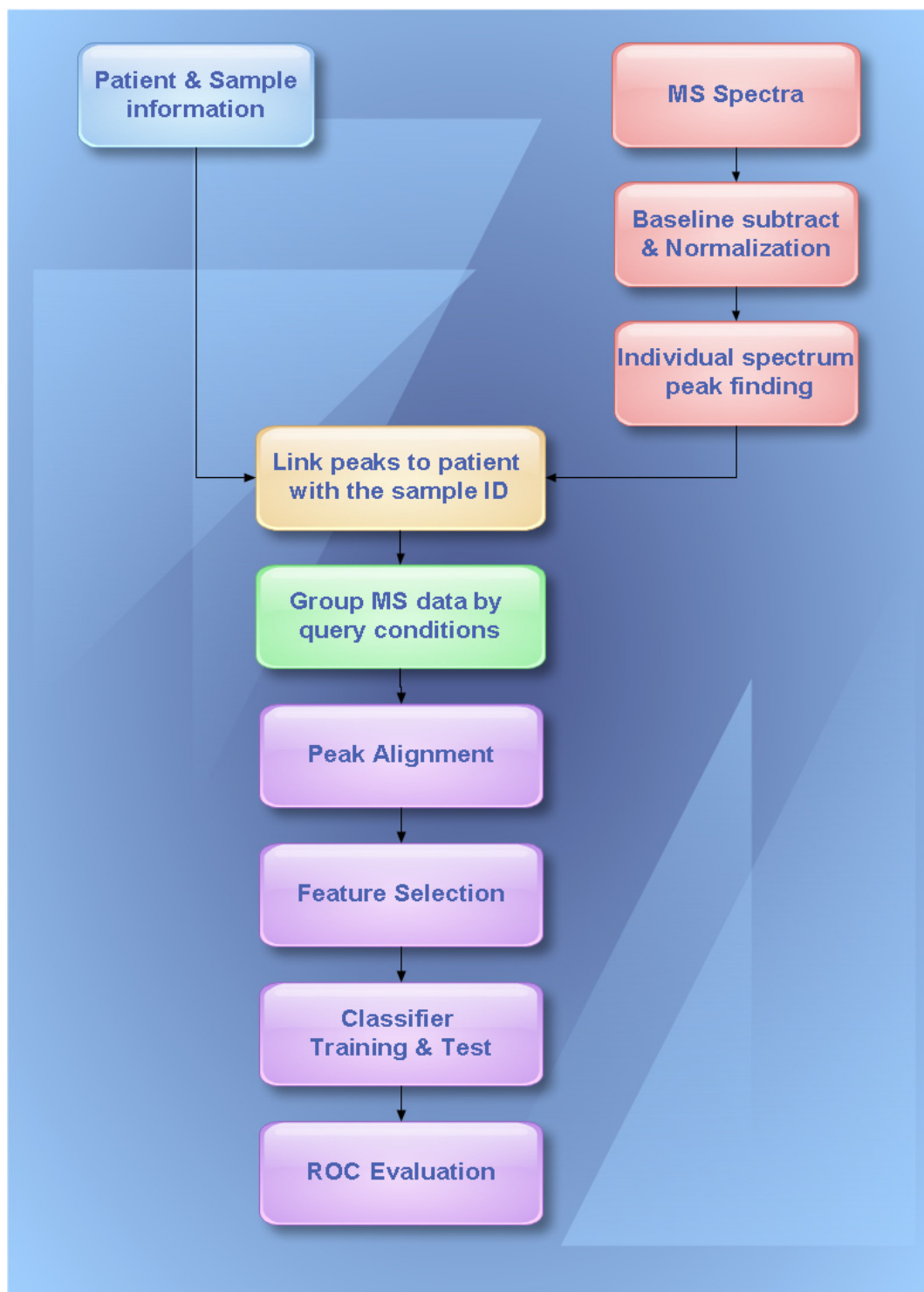


Figure 2
The workflow charts of the BioSunMS. To construct model for patients classification using peptide profiles generated by MALDI-TOF MS or SELDI-TOF MS, following steps are suggested. The first step is to input patients information, then preprocessing of MS spectra data for each patient, linking MS peaks to each patient through sample ID, peak alignment and creation of data matrix, t-test based feature selection and SVM-based model construction. Finally, the ROC curve was used to evaluate the performance of the model.

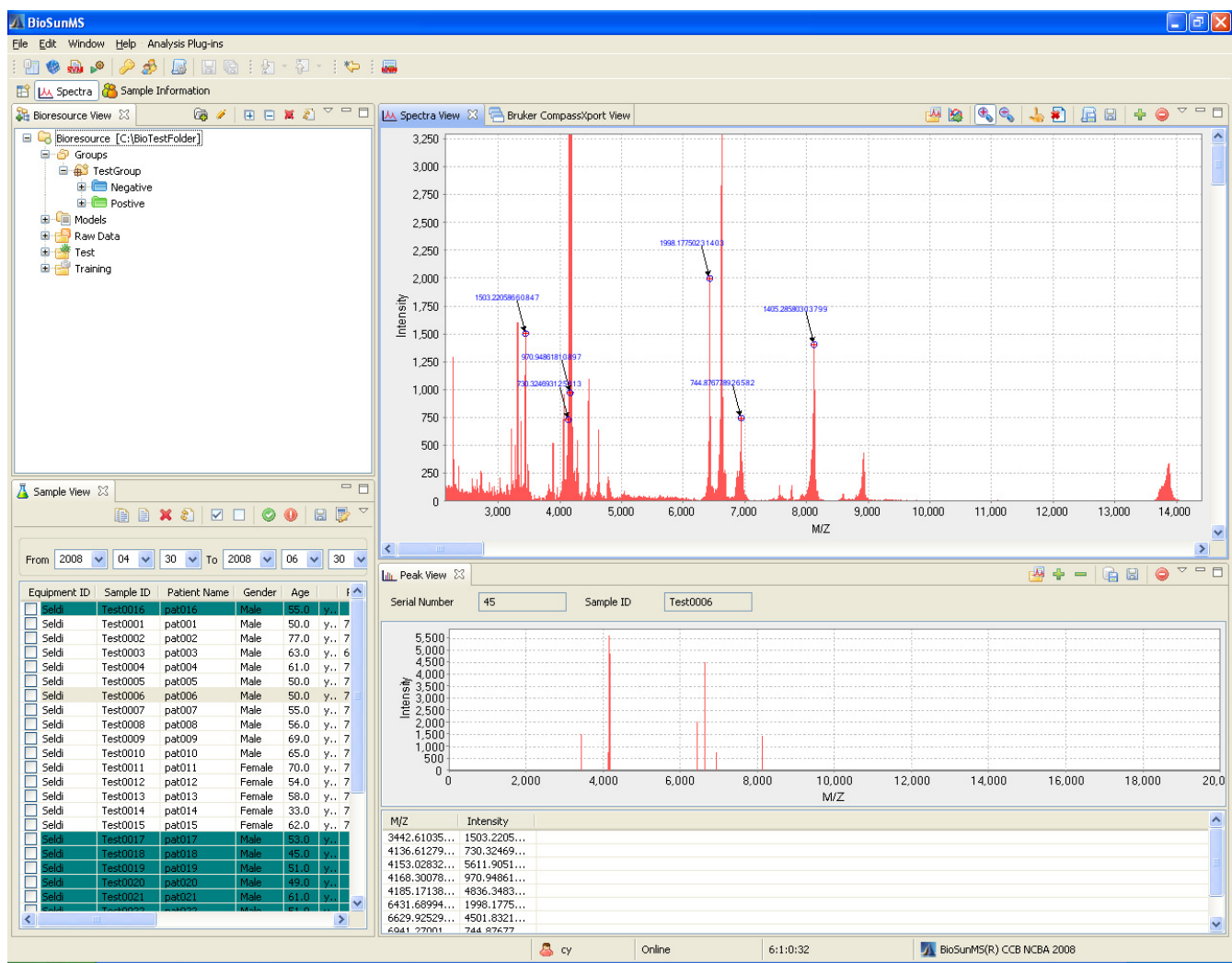


Figure 3
The interface for MS data display. The preprocessing of MS data for a particular patient was demonstrated, which included baseline subtraction, normalization, and automatic or manual peak extraction.

Therefore, data access and management can be password protected. The system always stores actions made by any users. The administrator can check the action history at any time.

From the User and Group Management View, BioSunMS administrator can add or edit the user and the group. If a user group has been permitted to access a kind of data, any users belonging to the group are permitted to read the data in the Sample View. Any user who doesn't belong to the user group will not be permitted to read the data in the Sample View. Any user can create a new record, and modify a record which belongs to the same user group. However, only administrators can create new users and user groups. By default a new record is accessible to all users.

To restrict access to a record, assign that record to the desired user group.

Sample and clinical information

Recording detailed information on collection, processing and storage of samples is crucial both for efficient reporting on biomedical study and for subsequent data analysis[25]. Many patients and sample related variables, such as gender, age, genetic factors, drug treatment, and symptom, dietary, and family history are particularly important in the context of clinical proteomics study. BioSunMS provides an efficient link between sample data and experiment results. Users can define the condition by selecting a series of fields to be used to automatically segregate spectra into groups. For example, in a typical profiling study,

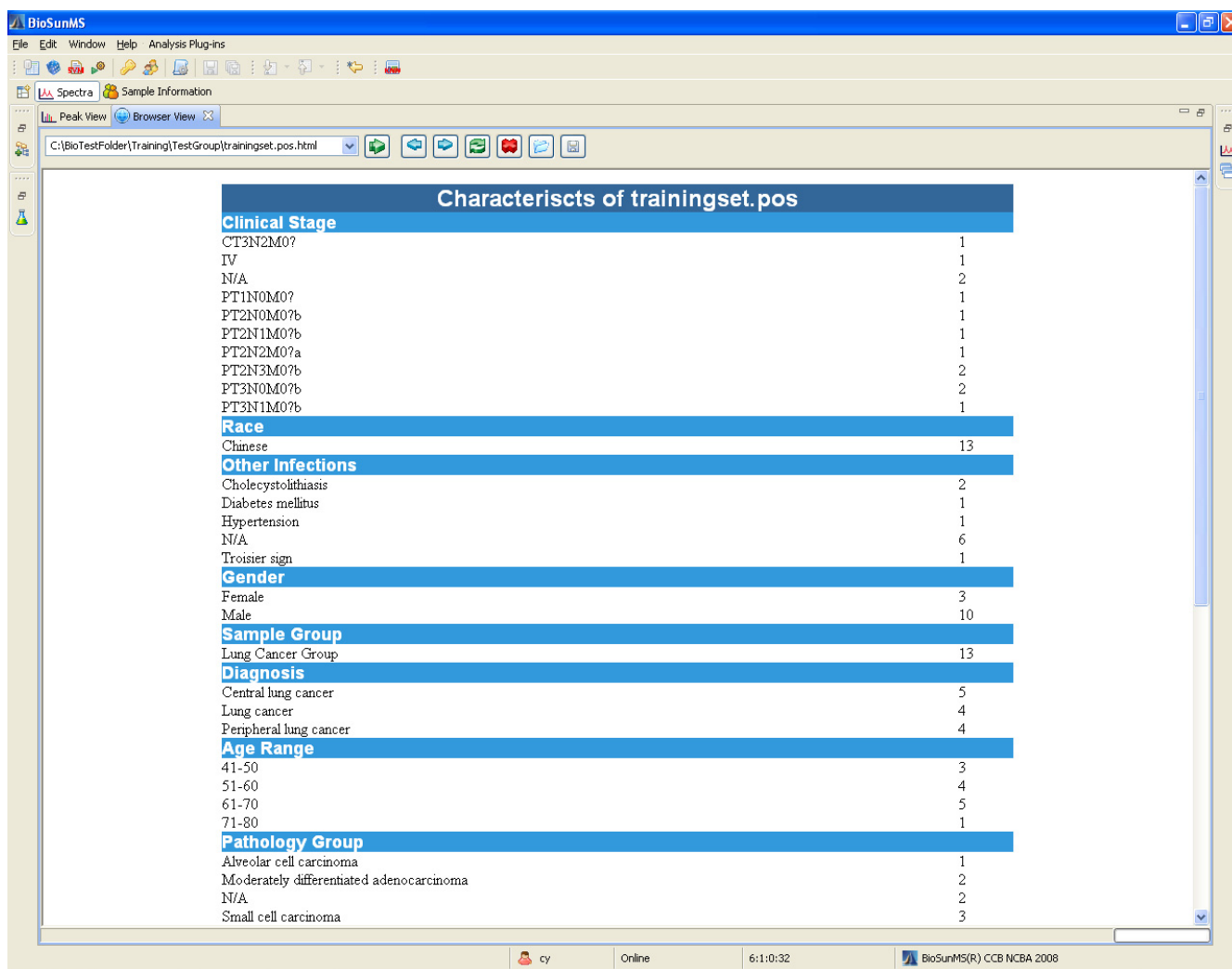


Figure 4
Summary of patients information. The patients characteristics were provided in a web page, which included clinical stage, age range, pathology diagnosis, and etc.

a condition might consist of the array type, sample type, different subjects (e.g. healthy and diseased), clinical laboratory result and diagnosis. After grouping the data by condition, BioSunMS will create a table with characteristics of the group of patients information (Figure 4). Finally, the data produced by the above steps can be exported for further processing outside of BioSunMS system using the export function provided by Bioresource View.

Pre-processing and visualization

BioSunMS is useful for the users with the need to store, process and analyze MALDI-TOF or SELDI-TOF MS data. It supports general mass spectrometry file format, such as mzML, mzXML, mzData and CSV. For mzML data format standard released by the HUPO-PSI and Institute for Sys-

tems Biology in June 2008, BioSunMS reads and writes it using a package of ProteomeCommons.org IO Framework [26] and ProteoWizard[27]. Currently, the majority of instruments store mass spectra in vendor specific formats. Users can convert raw file to universal file, for example, mzXML format using the converter programs provided by the vendors. At present, we have developed a plug-in which integrates CompassXport[28] for converting raw file of Bruker Corporation[29] to mzXML. Other plug-ins for format converting is under development.

BioSunMS provides basic visualization tools and advanced processing algorithms. The method and related procedure, BioConductor PROcess was used for baseline subtraction, smoothing, normalization and peaks identification. Our data visualization module aims to support

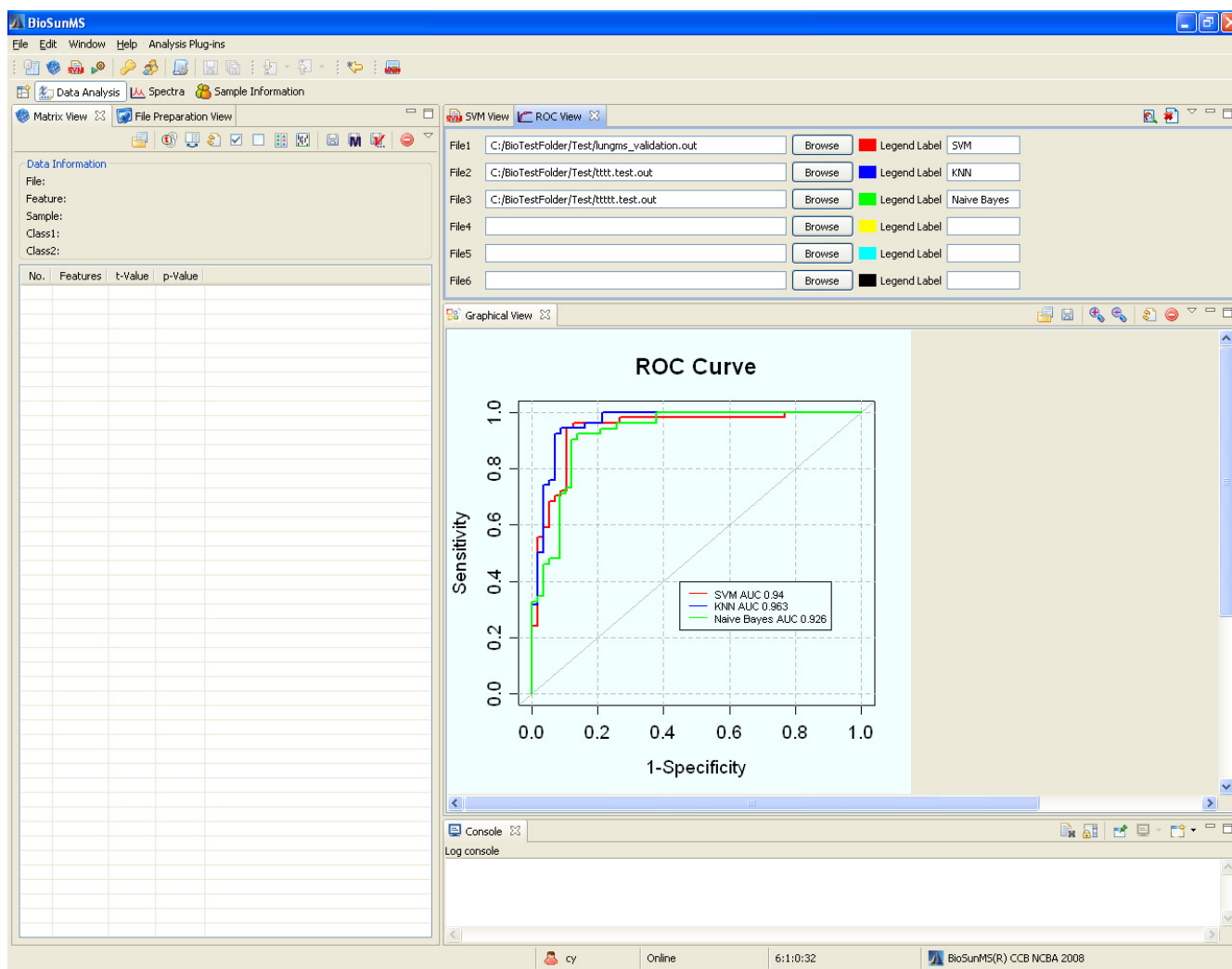


Figure 5
The screen shot of ROC curve. The ROC curve is used to evaluate the performance of model. The demonstrated example provides the ROC curves for classifiers constructed from SVM, KNN and Naive Bayes methods, respectively, and the correspondent areas under the ROC Curve are 0.94, 0.963 and 0.926.

analysts in finding interesting peaks in spectra, selecting them for further analysis and visualizing the selected peaks by automatic or manual detection.

Sample classification and sample prediction

The analysis of several spectra coming from biological samples belonging to different subjects (e.g. healthy and diseased) focuses on identifying discriminant values in spectra related to diseases. The BioSunMS allows users to select features using t-test, construct models using SVM, and predict new samples by the wizard with the default parameters. Users can compare the generalization performance of a range of classifiers by plotting their performance on the test set in ROC curve. BioSunMS also provides heat maps and hierarchical clustering. In the

future, we will gradually incorporate classification system Tclass [30] and sample class discovery system SamCluster [31] into BioSunMS, which were developed by our center for gene expression profile-based analysis.

R package

BioSunMS uses the R statistical programming language and some third part packages for processing and analysis of MS data. We implemented the communication between the BioSunMS and the R package with the class ncbaSpanker.System.R.RPackage. This class provides some methods for executing R scripts and commands. All communication with R language takes place via the file system. In general, BioSunMS writes a R script, makes the R system execute the script file, and output the results.

Then BioSunMS reads an R output file to retrieve the results. The RPackage class hides all the details of communication with R. So users with little knowledge about R language can accomplish their task easily. Because many programs for gene expression profile-based analysis have been provided in R package, and some algorithms can be directly used in peptide profiling analysis, we use R package as the background computation. In addition, many packages for MS data analysis and visualization are also provided, for example, GALGO[32], caMassClass [33] and so on. We will incorporate them into BioSunMS system in the future.

Development environment

Here we want to emphasize an important characteristic of BioSunMS. It can not only be used as the standard-alone program, but also can be a framework for developers. By taking full advantage of the editing and visualization components provided by Eclipse, the developers can focus entirely on the problems at hand. Moreover, BioSunMS was implemented in Eclipse platform, which made it more flexible and easier to adopt an algorithm as a plug-in. For example, we have successfully developed a plug-in, cn.org.biosun.knn, for sample classification using k-nearest neighbour (kNN) method.

Conclusion

BioSunMS is a plug-in based and flexible software for MS data management and analysis. It integrates patients information and MS data storage, process, sample classification and sample prediction into a single, user-friendly workbench. The project provides an alternative solution to analyze high throughput MS data from MALDI-TOF or SELDI-TOF.

The future for BioSunMS holds much potential with many plug-ins in development, including Web service, more statistical methods and machine learning algorithms, such as GA, Decision-Tree, Naïve Bayes method, and sample class discovery. There is also ongoing work for improving the plug-ins which allows researchers to enter patients data (i.e., past history, patient history, physical examination, social history, and family history data) in a structured way using a customizable domain model. Another major feature in development is online updates for plug-in from the BioSunMS update server. The current status of the BioSunMS development can be viewed at the BioSunMS website.

Availability and requirements

Project name: BioSunMS

Project home page: <http://ccb.bmi.ac.cn/biosunms/> or <http://sourceforge.net/projects/biosunms/>

Operating system(s): Windows, Linux

Programming language: Java, R, SQL

Other requirements: SUN JVM 1.5 or higher, R Package 2.5.1 or higher, MySQL 5.0 or higher

License: GNU GPL

Any restrictions to use by non-academics: Contact authors

Abbreviations

RCP: Rich Client Platform; SWT: Standard Widget Toolkit; MS: Mass Spectrometry; MALDI-TOF MS: Matrix-Assisted Laser Desorption/Ionization Time-Of-Flight Mass Spectrometry; SELDI-TOF MS: Surface-Enhanced Laser Desorption/Ionization Time-Of-Flight Mass Spectrometry; ROC: Receiver Operation Characteristic; SVM: Supported Vector Machines; kNN: k-Nearest Neighbour

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

YC designed and developed the system. XY, NW, AL and HW did the conceptual design of the system and gave some detailed technique supports. XZ and WL conceived the project, performed overall supervision and coordination. All authors read and approved the final manuscript.

Acknowledgements

This work was supported by Chinese Key Project for the Infectious Diseases 2008ZX10002-016.

References

1. Ruth E, Nicole U, Scott R, Martin M, Stephen S, Brian R, Jerald R, Garnet A, Leland H: **Early detection: The case for early detection.** *Nat Rev Cancer* 2003, **3**:243-252.
2. Petricoin EF, Belluco C, Araujo RP, Liotta LA: **The blood peptidome: a higher dimension of information content for cancer biomarker discovery.** *Nat Rev Cancer* 2006, **6**:961-967.
3. Petricoin EF, Zoon KC, Kohn EC, Barrett JC, Liotta LA: **Clinical proteomics: translating benchside promise into bedside reality.** *Nat Rev Drug Discov* 2002, **1**:683-695.
4. Agranoff D, Fernandez-Reyes D, Papadopoulos MC, Rojas SA, Herbst M, Loosemore A, Tarelli E, Sheldon J, Schwenk A, Pollok R, Rayner CF, Krishna S: **Identification of diagnostic markers for tuberculosis by proteomic fingerprinting of serum.** *Lancet* 2006, **368**:1012-1021.
5. Pan S, Zhang H, Rush J, Eng J, Zhang N, Patterson D, Comb MJ, Aebersold R: **High throughput proteome screening for biomarker detection.** *Mol Cell Proteomics* 2005, **4**:182-190.
6. Cannataro M: **Computational proteomics: management and analysis of proteomics data.** *Brief Bioinform* 2008, **9**:97-101.
7. Titulaer MK, Siccama I, Dekker LJ, van Rijswijk AL, Heeren RM, Sillevius Smitt PA, Luidert TM: **A database application for pre-processing, storage and comparison of mass spectra derived from patients and controls.** *BMC Bioinformatics* 2006, **7**:403.
8. Villanueva J, Philip J, Chaparro CA, Li Y, Toledo-Crow R, DeNoyer L, Fleisher M, Robbins RJ, Tempst P: **Correcting common errors in identifying cancer-specific serum peptide signatures.** *J Proteome Res* 2005, **4**:1060-1072.

9. **MathWorks** [<http://www.mathworks.com>]
10. **Agilent Technologies** [<http://www.chem.agilent.com>]
11. Sturm M, Bertsch A, Gröpl C, Hildebrandt A, Hussong R, Lange E, Pfeifer N, Schulz-Trieglaff O, Zerck A, Reinert K, Kohlbacher O: **OpenMS – an open-source software framework for mass spectrometry.** *BMC Bioinformatics* 2008, **9**:163.
12. Leptos KC, Sarracino DA, Jaffe JD, Krastins B, Church GM: **MapQuant: open-source software for large-scale protein quantification.** *Proteomics* 2006, **6**:1770-1782.
13. Hartler J, Thallinger GG, Stocker G, Sturn A, Burkard TR, Körner E, Rader R, Schmidt A, Mechtler K, Trajanoski Z: **MASPECTRAS: a platform for management and analysis of proteomics LC-MS/MS data.** *BMC Bioinformatics* 2007, **8**:197.
14. Li XJ, Yi EC, Kemp CJ, Zhang H, Aebersold R: **A software suite for the generation and comparison of peptide arrays from sets of data collected by liquid chromatography-mass spectrometry.** *Mol Cell Proteomics* 2005, **4**:1328-1340.
15. Bellew M, Coram M, Fitzgibbon M, Igra M, Randolph T, Wang P, May D, Eng J, Fang R, Lin C, Chen J, Goodlett D, Whiteaker J, Paulovich A, McIntosh M: **A suite of algorithms for the comprehensive analysis of complex protein mixtures using high-resolution LC-MS.** *Bioinformatics* 2006, **22**:1902-1909.
16. Katajamaa M, Miettinen J, Oresic M: **MZmine: toolbox for processing and visualization of mass spectrometry based molecular profile data.** *Bioinformatics* 2006, **22**:634-636.
17. **BioSun** [<http://www.biosun.org.cn/>]
18. **Eclipse.org Home** [<http://www.eclipse.org/>]
19. **Eclipse Rich Client Platform** [<http://www.eclipse.org/home/categories/rcp.php>]
20. **R Package** [<http://www.r-project.org/>]
21. **LIBSVM** [<http://www.csie.ntu.edu.tw/~cjlin/libsvm/>]
22. Jeff M, Jean-Michel L: *Eclipse Rich Client Platform: Designing, Coding, and Packaging Java Applications* Addison-Wesley Professional; 2005.
23. **Bioconductor PROcess** [<http://www.bioconductor.org/packages/bioc/1.8/html/PROcess.html>]
24. **JFreeChart** [<http://www.jfree.org/jfreechart/>]
25. Lyons-Weiler J: **Standards of excellence and open questions in cancer biomarker research: an informatics perspective.** *Cancer Informatics* 2005, **1**:1-7.
26. Falkner JA, Falkner JW, Andrews PC: **ProteomeCommons.org IO Framework: reading and writing multiple proteomics data formats.** *Bioinformatics* 2007, **23**:262-263.
27. Kessner D, Chambers M, Burke R, Agus D, Mallick P: **ProteoWizard: open source software for rapid proteomics tools development.** *Bioinformatics* 2008, **24**:2534-2536.
28. **CompassXport** [http://www.ionsource.com/functional_reviews/CompassXport/CompassXport.htm]
29. **Bruker Corporation** [<http://www.bruker.com/>]
30. Wuju L, Momiao X: **Tclass: tumor classification system based on gene expression profile.** *Bioinformatics* 2002, **18**:325-326.
31. Wuju L, Ming F, Momiao X: **Samcluster: An integrated scheme for automatic discovery of sample classes using gene expression profile.** *Bioinformatics* 2003, **19**:811-817.
32. Trevino V, Falciani F: **GALGO: an R package for multivariate variable selection using genetic algorithms.** *Bioinformatics* 2006, **22**:1154-6.
33. **caMassClass** [<http://cran.r-project.org/web/packages/caMassClass/index.html>]

Pre-publication history

The pre-publication history for this paper can be accessed here:

<http://www.biomedcentral.com/1472-6947/9/13/prepub>

Publish with **BioMed Central** and every scientist can read your work free of charge

"BioMed Central will be the most significant development for disseminating the results of biomedical research in our lifetime."

Sir Paul Nurse, Cancer Research UK

Your research papers will be:

- available free of charge to the entire biomedical community
- peer reviewed and published immediately upon acceptance
- cited in PubMed and archived on PubMed Central
- yours — you keep the copyright

Submit your manuscript here:
http://www.biomedcentral.com/info/publishing_adv.asp

