Spectroscopy 16 (2002) 81–87
IOS Press

A bioinformatics approach for mass spectrometry data processing: Applications to proteomics and small molecule analysis

Martin Sonderegger, Kristin Staniszewski, Andrew Meyers and Gary Siuzdak *
The Scripps Center for Mass Spectrometry and Department of Molecular Biology, The Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla, CA 92037, USA

Abstract. We have developed a web-based software system, *JULIAN*, that simplifies the process of relaying mass spectral information for chemists, protein chemists, biochemists and all others performing mass spectrometry experiments through a centralized mass spectrometry laboratory. *JULIAN* allows for relative ease in submitting compound information as well as instant access to analysis results from any networked computer equipped with a web browser. Compound information is centralized in a Microsoft Access database and results are available in Adobe's portable document format (PDF) from an NT4 server. This gives researchers the ability to easily obtain data and allows the analysts in the mass spectrometry lab to browse analysis results when assisting researchers with their inquiries. Due to this web-based design *JULIAN* is independent of the mass spectrometers' hardware and operating system. Approximately seven hundred on-site and off-site users have utilized *JULIAN* transmitting over 40,000 analyses. The conversion from paper to electronic mass spectrometry data processing has enabled our Center to receive compound information, perform analysis, and relay the results four times faster than required previously.

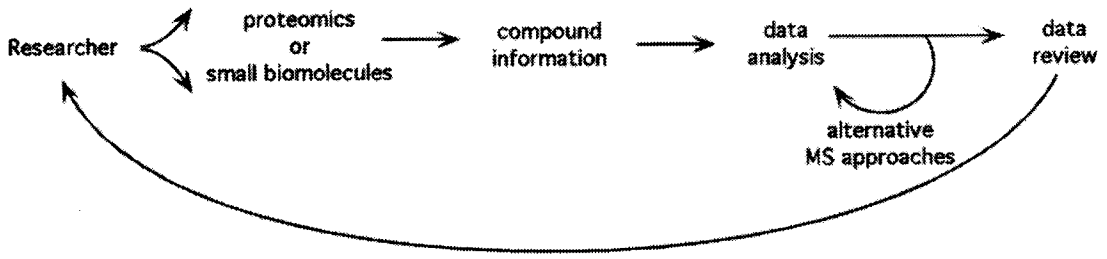
Keywords: Data processing, bioinformatics, programming, proteomics, small molecules

1. Introduction

Mass spectrometry has the capacity to provide valuable information in such diverse areas as drug discovery, biomolecule characterization and proteomics. The generation of large volumes of data poses the challenge of processing the information into a meaningful and practical form [1–5], and relaying the information rapidly to the researchers. To address this challenge we have developed an electronic system that simplifies and expedites the process for both the researchers and those who perform the mass spectrometry experiments. Our goal has been to use a web-based system as the sole means of relaying information. *JULIAN*, the electronic "paperless" system that we have developed allows for quicker access to the submission procedure, instant access to analysis results and easier retrieval and interpretation of results.

From its inception in 2000, nearly 700 on-site and off-site users have utilized the paperless sample system employed at the Center of Mass Spectrometry at the Scripps Research Institute, transmitting over 40,000 analyses. With *JULIAN*, users have the ability to quickly and effortlessly submit compound information in html-form, via the Internet, to the mass spectrometry laboratories. Users may elect to have their compounds analyzed by one or more of a variety of techniques including: Matrix-assisted laser desorption/ionization time-of-flight reflection mass analysis (MALDI-TOF); matrix-assisted laser desorption/ionization Fourier transform-ion cyclotron resonance mass analysis (MALDI-FTMS); electrospray

*Corresponding author. E-mail: siuzdak@scripps.edu.



Scheme 1. Overview of the compound analysis process. A researcher submits proteomic or small biomolecule compound information to the mass spectrometry lab. Once the compound has been analyzed the data is reviewed and results are relayed to the researcher.

	Method	Further Information
Exact mass:	>> MALDI-FTMS	o used for high resolution/accuracy o typically < 1500 Da
Unit mass:	>> MALDI-TOF	o used for large molecules and complex mixtures o 1000 - 200000 Da
	>> Electrospray	o typically 100 - 3000 Da o proteins also possible (typically < 40000 Da)
	>> GC/MS	o small hydrophobic molecules o < 500 Da
	>> Try all possible approaches	o We try as many methods as necessary and possible to get unit mass on your sample
Combinations:	>> Choice of more than one method on next page	o You may choose the methods you want us to do

Fig. 1. Depending on desired resolution and the size and complexity of the compound, researchers may elect to have their compounds analyzed by one or a combination of methods such as MALDI-FTMS, MALDI-TOF, electrospray and GC/MS.

ionization (ESI) and/or gas chromatography mass spectrometry (GC/MS) (Fig. 1). By implementing *JULIAN*, the lab has the capacity to process thousands compounds a month. The progression from paper to electronic mass spectrometry has enabled our lab to receive compound information, analyze the compound and relay the analysis results to the researcher in a fraction of time required previously.

2. Overview of the paperless system design

A web-based design was created to facilitate a paperless data processing system for chemists, protein chemists, biochemists and all others performing mass spectrometry experiments through a centralized mass spectrometry laboratory. Our approach overcomes challenges encountered with paper and other less thorough computer-based systems. Furthermore *JULIAN* is independent of the mass spectrometers' hardware and operating systems, allowing for the retrieval of compound information across different platforms. Compound information and analysis results are centralized in a Microsoft *Access* database located on an NT4 server. This gives researchers the ability to easily obtain their analysis results in Adobe portable document format (PDF), and allows the analysts in the lab to browse analysis results when assisting researchers with their inquiries. With this electronic method of data distribution, researchers

have the ability to submit compound information from any networked computer equipped with a web browser. No paper forms are needed.

3. Details

The web-based software can be broken down into five distinct sections: (1) information submission, (2) mass analysis, (3) data review and (4) automated email notification upon analysis completion, coupled with instructions to assist researchers when downloading their analysis results from the website. (5) An additional section includes administrative tools such as database maintenance, statistics and accounting.

For security and accounting purposes every user is required to log onto the system. They then choose the general category for analysis: proteomics or small molecules. New users are required to submit information such as email, name and phone number; for future reference, this information is stored in a database. The scientist then submits all necessary information pertaining to the compound in a web form (Fig. 2). For small molecules this information includes molecular formula and mass, solution concentration and solvent, as well as compound purity and toxicity. In addition scan range, compound type, functional groups and general comments can be submitted to assist the analyst when running their compounds. For proteomics the web form requires the submission of the molecular weight and protein quantity, purity and concentration. Biological source, sequence, possible modifications, storage instructions and comments are also stored in the database.

Analysts in the mass spectrometry laboratory can also access the stored information from a web browser where compounds are listed according to the analytical technique requested (Fig. 3). The cover page of an automatically prepared PDF file contains all the information about the sample as well as space to report analysis results (Fig. 4). Some embedded Java-scripts facilitate the automatic calculation of the expected masses and the error on the observed mass. Mass spectral data are printed to a PDF file and attached to the cover page. If necessary for further analysis, or when requested by a researcher, a compound can be forwarded to another method of analysis.

After a final data review – which can be performed from any web browser – the data is stored on the server for retrieval and the scientist is notified by email about the completion of the analysis. The file containing all the compound information as well as mass spectral data is only about 100 kB in size. Typically our system allows online access to the data for three months before it is permanently stored on a CD.

In addition to improving efficiency, the paperless system has implemented some other tools to make it easy for the user, those performing mass analyses and those interested in throughput. As all the information is stored at one central location, generating statistics about the performed tasks is easily accomplished. Detailed lists to track each laboratory as well as each user's sample load can be generated. A feature to send email messages to the users for the paperless system has also been implemented to address just researchers interested in a single method of analysis or to reach everyone who submits compounds to the facility. There has also been care taken to provide an online administration interface to add or update user information and to change incorrectly submitted compound details. Submissions may be deleted from the database completely or just removed temporarily if a researcher decides to send the compound at a later time.

In order to make a system such as *JULIAN* functional, there are a few necessary tools. The basis is a Microsoft NT4 Server installed on a PC system (Intel PIII 550 MHz, 256 MB RAM, 2 mirrored 13 GB hard drives) which provides the Microsoft Internet Information Server (MS IIS 4). User and compound

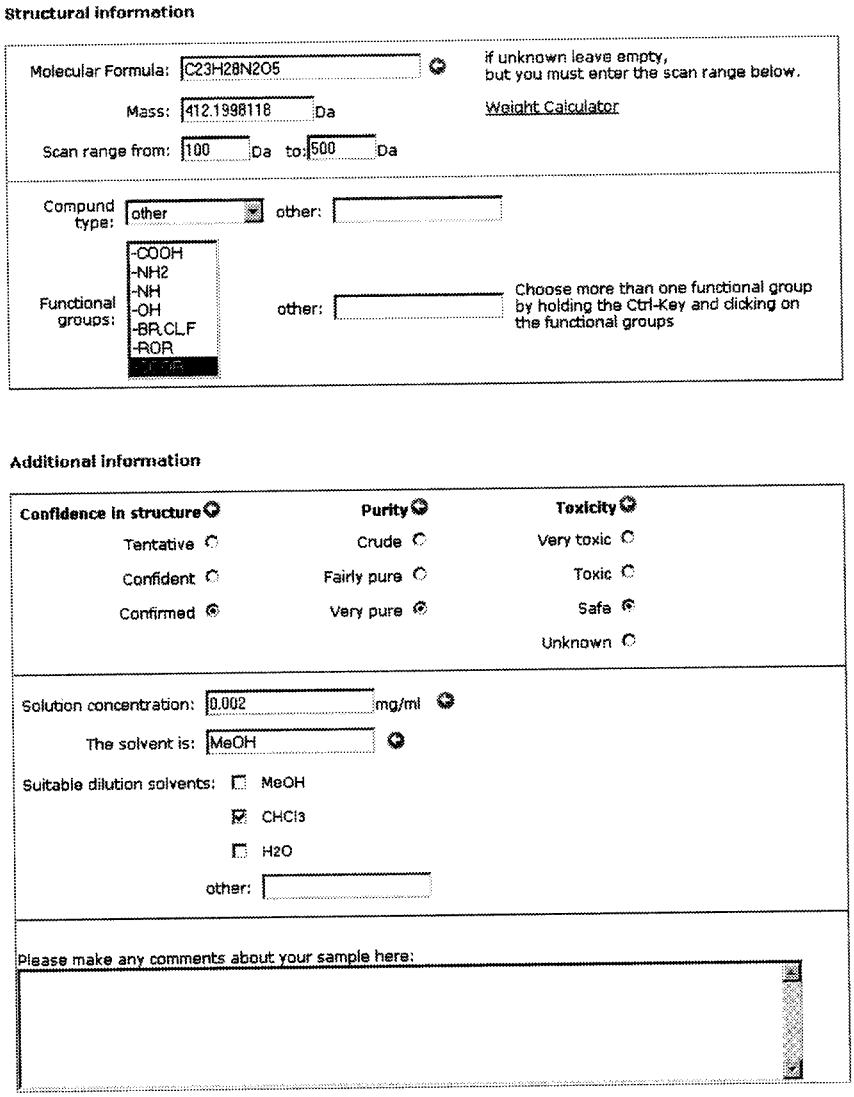


Fig. 2. All compound structural information including molecular formula and mass, compound type and functional groups is submitted on a web form. Additional information such as purity, toxicity and solution concentration is also contained in this form.

information is stored in a Microsoft *Access* Database. *Access* has the capacity to handle hundreds of samples a day, yet is straightforward enough to convert the system to MS SQL Server or to a similar database application. A majority of the code that drives the paperless system is embedded in Active Server Pages. Active Server Pages make it easy to create dynamic web pages, query databases, deal with user input through web forms, and keep track of a user's session after logging in. Some maintenance – such as updating the statistics – is accomplished using VB script and the scheduler of Microsoft NT. The results from each mass spectrometer are printed to a pdf file, which requires Adobe Acrobat to be installed on each instrument's computer. During the analysis process the PDF files are stored in a shared folder on the server which is accessible from PC and Macintosh computers through the local area network. After a final data review, ASP scripts move the files to a folder in the file system which is accessible to researchers.

Sample ID	Formula	Mass (Scan Range)	Exact Mass	MH+	MNa+	Concentration in solvent	Suitable Solvent			
							MeOH	CHCl3	H2O	other
RSBS192	C22H22N2O5	394	394.1529	395.1601	417.1421	1 mg/ml in CHCl3		x		
RSWS192	C161H184ClN21O35S2Si2	3126.2	3126.1932	3127.2004	3149.1824	0.05 mg/ml in MeOH	x			EtOH
RSSW192	C22H37NO5Si	423.2440872	423.2441	424.2514	446.2333	0.1 mg/ml in Ether		x		
KCAF5204	C44H83NO6S2Si3	869.4970	869.4969	870.5042	892.4862	2 mg/ml in CHCl3		x		
SNOV5205	C24H32O5	400.515	400.225	401.2322	423.2142	0.1 mg/ml in MeOH	x	x		
LCAF5209	C27H41NO5S2	523.2426	523.2426	524.2499	546.2318	2 mg/ml in CHCl3		x		
SNL55213	C19H30O4	322.21 150-600	322.2144	323.2217	345.2036	5 mg/ml in CDCl3		x		
S4MS5213	C16H22N4O3	318.2	318.1692	319.1765	341.1584	0.01 mg/ml in MeOH	x			
SCLW5217	C23H28N2O5	412.1998118 100-500	412.1998	413.2071	435.189	0.002 mg/ml in MeOH		x		

Fig. 3. Compounds are listed according to analysis technique requested. The samples shown here have been submitted for electrospray analysis. Sample ID, formula, scan range, concentration and suitable solvents are given. Exact mass, mass of the protonated compound and mass of sodium abducts are provided as well.

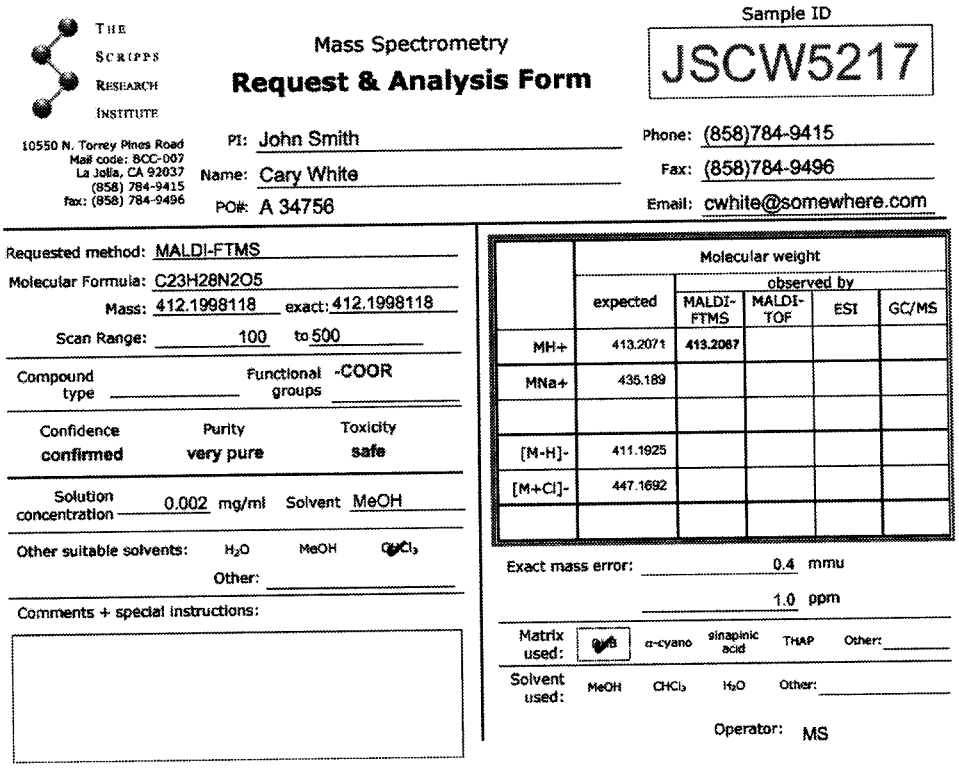
Simultaneously, an email is automatically sent to the researcher using Collaboration Data Objects for Windows NT Server (CDONTS) over SMTP. Overall *JULIAN* meets many of the needs for users requiring rapid analyses capabilities as well as access to their data. The primary proof of its reliability is that it has been successfully operational for over two years.

4. Future considerations

Automated proteomics is currently being integrated into *JULIAN*. We define automated proteomics as the "handless" processing of proteins with robotics followed by automated MS analysis and searching. No significant changes are required of the submission and data analysis and review aspects of the system to integrate this branch of analysis. The remaining step, transferring data to the researcher, requires the greatest modification with challenges associated with transferring very large data files, numerous search outcomes, spectra and coverage maps remains. As with the *JULIAN* system, information must be transferred in a user-friendly means minimizing user input and maximizing automated data manipulation. We project with the *JULIAN* system in place it would be feasible to perform 1000's of protein identifications and return the results to the researcher in a twenty-four-hour period. However, the trypsin digest period (15 hrs) associated with each set of samples effectively limits the maximum number of IDs to 192 per day. We are currently experimenting with the idea of clone systems. In this setup, the *JULIAN* system acts as the data pipeline of which the protein ID data packet containing the search data and associated spectra is transferred. The researcher in turn has an identical computer system to the high throughput system. This allows the research to analyze the data or view the pre-processed data independently from the high throughput system.

An ongoing effort is also underway to generate a searchable database for human natural products (HNP's). This extension of *JULIAN* will allow for searches of formula, elemental composition, mass, and tandem mass spectrometry data. Searches can be performed online and, as with the current version of *JULIAN*, this application provides web-based retrieval of information.

(a)



THE SCRIPPS RESEARCH INSTITUTE

10550 N. Torrey Pines Road
Mail code: BCC-007
La Jolla, CA 92037
(858) 784-9415
fax: (858) 784-9496

Mass Spectrometry

Request & Analysis Form

Sample ID: JSCW5217

PI: John Smith
Name: Cary White
PO#: A 34756

Phone: (858)784-9415
Fax: (858)784-9496
Email: cwhite@somewhere.com

Requested method: MALDI-FTMS
Molecular Formula: C23H28N2O5
Mass: 412.1998118 exact: 412.1998118
Scan Range: 100 to 500

Compound type: _____ Functional groups: -COOR

Confidence: confirmed Purity: very pure Toxicity: safe

Solution concentration: 0.002 mg/ml Solvent: MeOH

Other suitable solvents: H_2O MeOH $CHCl_3$
Other: _____

Comments + special instructions:

	Molecular weight				
	expected	observed by			
		MALDI-FTMS	MALDI-TOF	ESI	GC/MS
MH+	413.2071	413.2067			
MNa+	435.189				
[M-H]-	411.1925				
[M+Cl]-	447.1692				

Exact mass error: 0.4 mmu
1.0 ppm

Matrix used: DHB α-cyano sinapinic acid THAP Other: _____

Solvent used: MeOH $CHCl_3$ H_2O Other: _____

Operator: MS

(b)

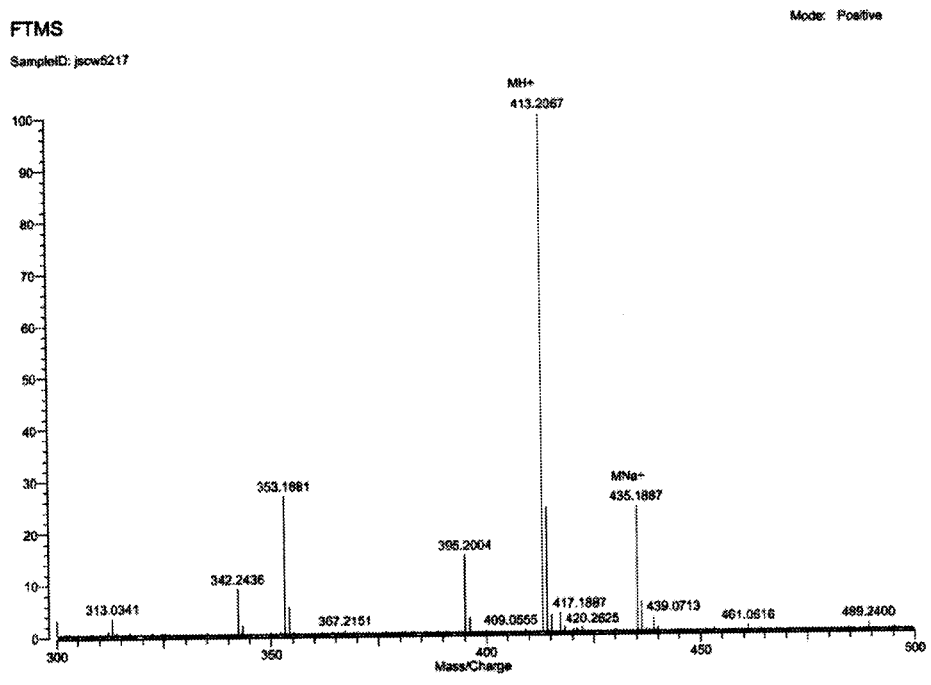


Fig. 4. The spectral data is attached to a request and analysis form cover page. This cover page contains the researchers contact information, a summary of the sample information and results, as well as space to provide special instructions and comments.

References

[1] J.P. Kurvinen, P. Rua, O. Sjovall and H. Kallio, Software (MSPECTRA) for automatic interpretation of triacylglycerol molecular mass distribution spectra and collision induced dissociation product ion spectra obtained by ammonia negative ion chemical ionization mass spectrometry, *Rapid Communications in Mass Spectrometry* **15** (2001), 1084–1091.

[2] W. Pusch, K.O. Kraeuter, T. Froehlich, Y. Stalgies and M. Kostrzewa, Genotools SNP MANAGER: a new software for automated high-throughput MALDI-TOF mass spectrometry SNP genotyping, *Biotechniques* **30**(1) (2001), 210–215.

[3] K.J. Hart, Software development for mass spectrometric analysis, *Rapid Communications in Mass Spectrometry* **10**(3) (1996), 393–398.

[4] H.I. Field, D. Fenyo and R.C. Beavis, RADARS, a bioinformatics solution that automates proteome mass spectral analysis, optimises protein identification, and archives data in a relational database, *Proteomics* **2**(1) (2002), 36–47.

[5] P.J.F. Watkins, I. Jardine and J.X.G. Zhou, Mass spectrometry software for biochemical analysis in electrospray and fast atom bombardment modes, *Biochemical Society Transactions* **19**(4) (1991), 957–962.