Proceedings of the 26th Annual International Conference of the IEEE EMBS
San Francisco, CA, USA • September 1-5, 2004

Analysis and Robot Pipelined Automation for SELDI-TOF Mass Spectrometry

Gil Alterovitz[1], Manuel Aivado[2], Dimitrios Spentzos[2], Towia A. Libermann[2], Marco Ramoni[3], and Isaac S. Kohane[3]

[1] Health Science and Technology/Electrical Engineering & Computer Science, Massachusetts Institute of Technology, Cambridge, MA, USA. [2] Genomics Center of the Beth Israel Deaconess Medical Center, Harvard Institutes of Medicine, Boston, MA, USA. [3] Harvard Medical School and Harvard Partners Center for Genetics & Genomics, Boston, MA, USA.

***Abstract*—Surface-enhanced laser desorption/ionization time-of-flight mass spectrometry (SELDI or SELDI-TOF MS) with protein arrays has facilitated the discovery of disease-specific protein profiles in serum. As array technologies in bioinformatics and proteomics multiply the quantity of data being generated, more automated hardware and computational methods will become necessary in order to keep up. Robot Automated Sample Preparation and Analysis Pipeline for Proteomics (Raspap) in SELDI provides a solution from the lab bench to the desktop. In this approach, the entire processing of protein arrays is delegated to a robotics system and the Bioinformatics Automated Pipeline (BAP) performs data mining after SELDI analysis. A key part of BAP is the creation of a journal-styled report in HTML (with text, embedded figures, and references) which can be automatically emailed back to the engineers/scientists for review. An object-oriented tree-based structure allows for the derivation of conclusions about the data and comparison of multiple analyses within the generated report.**

Testing yielded improvement in the resulting assay coefficients of variation (CV) from 45.1% (when done manually) to 27.8% (P <0.001). A large biological dataset was also examined with the Raspap approach and consequent results are discussed.

***Index Terms*—Automation, Bioinformatics, Biomedical Laboratories, Mass Spectroscopy, Proteins, Protocols, SELDI**

I. Introduction

SELDI time-of-flight mass spectrometry has been developed for rapid peptide and protein profiling of complex biological samples [1]. In SELDI, samples can be incubated on protein array surfaces having different properties (e.g. hydrophobic, ionic, and metal affinity). Depending on the array's binding properties, different peptides and proteins will be retained on its surface. After dispensing a matrix molecule onto the sample spots to aid in the desorption process, a SELDI time-of-flight mass spectrometer can be used to generate a spectrum of peaks.

Pioneering work by the Liotta lab [2] applied protein arrays to proteomic profiling. While still in its infancy, the growth in this new field suggests that more advanced techniques will be needed to deal with larger proteomic sets both in terms of hardware and software automation.

Automation has been used in laboratory robotics previously [3, 4]. Additional automation technologies have pursued creating specialized protein arrays using novel methods [5]. However, this is the first work which includes automation both in hardware to prepare samples and in software to automatically do analysis to support SELDI.

If done manually, it is very time consuming to do replicates, expensive, and technically challenging to maintain low variation between spots. Thus, this paper assesses the potential usefulness of replicating sample spots by looking at intra-assay variability. This also serves as a second step in exploring how automation can be useful in increasing the number arrays that can be processed- thus allowing for replicates to become more practical.

Early work in automation for mass spec focused on identification and study of individual molecular peaks. With current array technology, the emphasis has shifted toward a many-to-one mapping of a multitude of proteins (i.e. a protein profile fingerprint) to a diagnosis state. While identification of individual protein peaks can be important for biological validation, the process is time consuming and is usually done only after computational analysis of the proteomic profiles. The analysis portion of our system, BAP, was designed with both of these issues in mind. The key point of BAP (and the entire Raspap system) is to get answers back to the scientist/engineer in a human-readable form (e.g. a journal-styled report) as soon as possible.

BAP represents the first automated pipeline that incorporates a full technical reporting engine to effectively close the loop of data analysis. It provides both a presentation and quick interpretation of the results for a variety of data analysis methods.

II. Materials and Methods

An overview of the Raspap system is shown in Fig. 1 and is explained in more detail in subsequent sections of this paper. Briefly, the procedure is as follows. The robot platform must first be loaded. In this setup, all cells are empty except the following: TL1, P1, and P3 have tips; P5 and P11 have waste containers; P6 is the bioprocessor cell; P7 is water; P9 has serum samples; P10 has wash mixture. The Raspap robot protocol is then loaded from a computer and run on the robot. Next, SELDI is performed to obtain

Manuscript received April 10, 2004 This work was supported in part by National Institutes of Health Grants U24 DK 58739 and 1RO1 CA 85467, the Whitaker Foundation, and the German Dr.-Mildred-Scheel Cancer Foundation.

the protein profile peaks. These peaks are analyzed by BAP and a report is generated for further interpretation. Based on the results, subsequent biological work may be done to identify candidate proteins recommended by BAP based on their importance in predicting a disease state. Alternatively, the feedback can be used to design further experiments.

A. Initial Serum and Plasma Samples

Blood from 80 people was collected. In addition, plasma samples from twelve volunteers were obtained from a plasma bank.

B. Setting up the Protein Arrays

If not otherwise indicated, the protein arrays (Ciphergen) were processed in accordance to the protocol listed on the Raspap web site [6] using a fully automated liquid-handling robotic system (Biomek FX™, Beckman Coulter, Fullerton, CA, USA) equipped with a 96-channel 200 µl head. H50 arrays were used for hydrophobic interaction chromatography and processed accordingly [6]. Immobilized Metal Affinity Capture (IMAC3) and Cation Exchange Chromatography (CM10) protein arrays can be processed in a similar manner to the H50 arrays except for some slight modifications to charging (for IMAC3), washing (both), and activating (for CM10) the surfaces. Sinapinic acid (SPA) was used as here as the matrix molecule to help in the subsequent laser desorption. For the manual experiments, washing and incubating steps were performed on a titer plate shaker (Lab-Line Instruments, Melrose Park, IL, USA) using the identical times to the automated procedure.

C. SELDI Time-of-Flight Mass Spectrometer

Mass spectrometry was performed with the Protein Biology System II SELDI-TOF MS reader (Ciphergen). The reader was externally calibrated every day with 10 different calibrants. Additionally, prior to each series of experiments, the spectrometer's resolution and sensitivity were examined to ensure consistent performance.

The mass spectra were derived from 80 laser shots per spot, collected at two different laser settings, as recommended by the manufacturer. The low-energy protocol allowed for detection of peptides and proteins < 10,000 Da and the high-energy protocol was used for capturing proteins between 10,000 to 40,000 Da

D. Bioinformatics Analysis Pipeline

BAP was developed as a result of consultation with researchers in the biomedical sciences and bioinformatics at several Harvard-affiliated centers including Harvard Partners Center for Genetics and Genomics (i.e. the bioinformatics center for Massachusetts General Hospital and Brigham and Womens Hospital), Children's HST Informatics Program (Children's Hospital), and Genomics

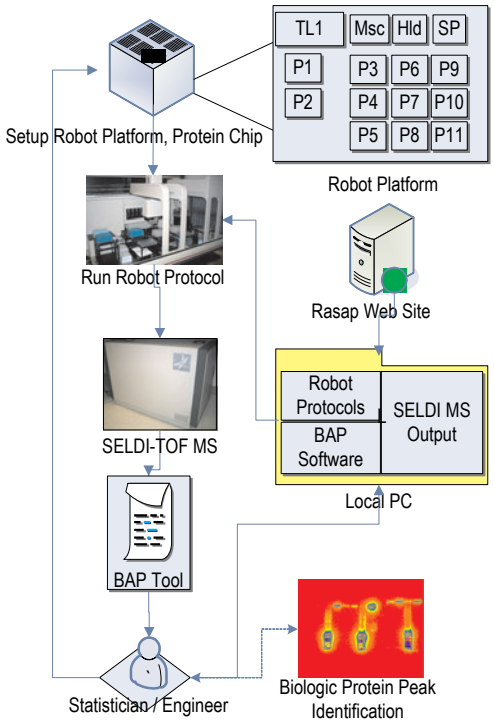

Fig. 1. Raspap System Overview

Center of the Beth Israel Deaconess Medical Center (Beth Israel Deaconess Medical Center). Through this collective experience, it was discovered that the biologist collaborations with the computational faculty could be abstracted into a schema. In practice, the BAP tool has helped to make for fewer, yet more productive meetings. It can serve as a first-line of analysis so biologists get prompt insight into potential findings and pitfalls. At the same time, it allows computational faculty to work on developing novel methods (which can then be encapsulated via Matlab or Java-compatible code).

With this framework in mind, BAP was developed so that the biologist can submit data directly to a central system (as shown in Fig. 2) via a front-end web server that deposits files for BAP to analyze. Alternatively, BAP can reside on the same computer that controls the SELDI machine in Raspap- allowing for seamless automation locally. Data is forwarded to the core computational engine. This engine uses internal (e.g. support vector machines) algorithms, global algorithms (e.g. k-fold cross validation), report specifications, and decision rules encoded in core modules. The computational engine processes the data through the analytical pipeline. The report generator then produces a final report based on outputs encoded in the resulting object-oriented tree structure. Afterwards, a journal-style report is emailed to the biologist and/or statistician/engineer for review as well as saved on the network/local file server in HTML format for later retrieval. These methods were

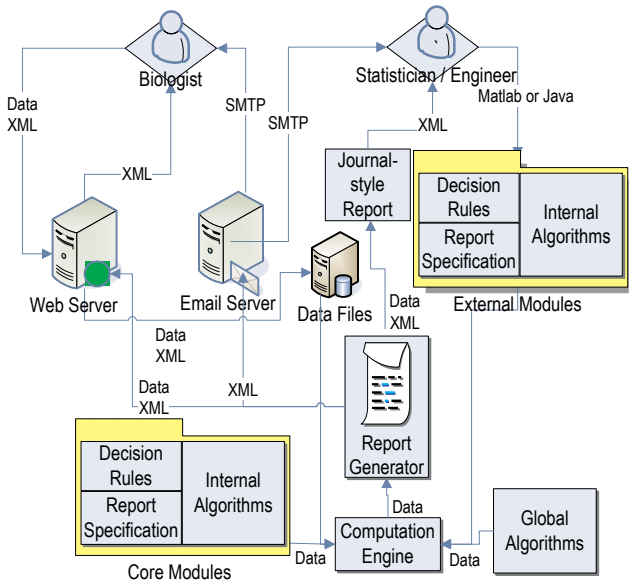

Fig. 2. Bioinformatics Analysis Pipeline Overview

implemented in Matlab and Java.

The data analysis methods and results are encoded within an object-oriented tree structure. The root node is the main 'document.' This is followed by sections of the final report such as 'Methods,' 'Results,' and 'References.' Under each section is a method category. Current method categories in the pipeline do basic statistics, statistical analysis, supervised learning, and unsupervised learning. A small sampling within the basic statistics and analysis include: kurtosis, skew, normality test, and box-whisker plots. Unsupervised learning includes methods like PCA [7] and clustering (e.g. hierarchical and k-means [8]). The supervised learning methods are support vector machines (SVM), k-nearest neighbors, naïve Bayesian classifier (NBC) [9], greedy NBC-based feature selection, and logistic regression. Under the methods are a series of properties that relate to the analysis, decision rules (used to process and interpret the results), as well as report-specific information (e.g. figures to generate, etc.).

The decision rules in the object-oriented tree can dynamically control the interpretation of the results. For example, after the calculation of performance metrics of various supervised classification methods is complete, the most *sensitive* test (or other pertinent variable) can be determined and reported automatically. Similarly, explanations of particular methods and their results can be encapsulated so that they are only reported if the dataset requires them.

E. Raspap Applied to Large Dataset

Raspap was used to analyze serum from 74 patients with cancer [6], 39 control patients, and 24 healthy persons at

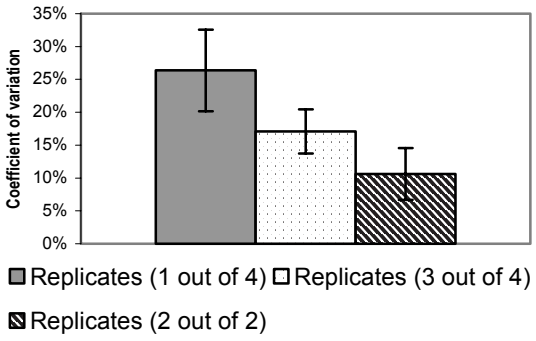

Fig. 3. Replicate variation. The peaks were defined by an SNR>2 in 1 out of 4 replicates, 3 out of 4, or 2 out of 2 replicates. The error bars represent standard deviation.

Harvard Medical School (USA) and University of Düsseldorf (Germany). The serum was processed with anion exchange chromatography and fractions of pH 5 and pH 9 were run on CM10 arrays. Both an organic fraction and unfractionated serum were run on H50 arrays.

III. Results

A. Automation and Replicate Variation

Automation of the protein chip experiments was examined to see if it improves the reliability of protein profiles. Serum samples were loaded in duplicates on H50 ProteinChip arrays: 20 manually, and 80 using an automated robot system. From the resulting SELDI mass spectra, peaks with an SNR $\geq$ 2 were utilized to determine the average CV for each pair of duplicates. The overall average CV was 45.1% $\pm$ 43.2% and 27.8% $\pm$ 30.9% for the manual and the automated procedure, respectively. This advantage for the automated procedure was statistically significant with P <0.001 according to both Mann-Whitney and ANOVA-tests. Therefore, all following experiments were carried out using the robotic system.

Replicate spot variation was examined for intra-assay variation. The average CVs between replicate spots were calculated for twelve samples using all of the peaks between 2,500-40,000 Da with an SNR $\geq$ 2 in at least one out of 4 replicates. The result was an intra-assay CV of 26.4% $\pm$ 6.2% with an average of 110.3 $\pm$ 5.6 peak positions used per sample. This CV was consistent with the one obtained from the previously discussed analysis on automation. Using the same samples, but assessing only peaks with an SNR $\geq$ 2 in at least 3 of 4 replicates, the intra-assay variation reduced to 17.1% $\pm$ 3.3% and the number of interrogated positions per sample decreased to 73.2 $\pm$ 4.2 peaks. Using only two replicates and peaks with an SNR $\geq$ 2 in both replicates led to an intra-assay variation of 10.6% $\pm$ 3.9% where the examined peaks per sample was 69.8 $\pm$ 4.0 (see Fig. 3).

B. *Results from Large Dataset*

A subset of the machine learning method results are shown in Table I regarding the cancer dataset used for this project. The most accurate predictor was SVM while the naïve Bayesian classifier was the most specific and sensitive. This is slightly better than a widely employed protein metric (used as a proxy for another disease, namely prostate cancer). In prostate cancer, prostate specific antigen (PSA) has the following characteristics (for PSA > 10.0 ng/ml [10]): 65.4% accuracy, 82.0% specificity, 41.6% sensitivity.

Table I. Classification results for using three machine learning methods.

CLASSIFIER	ACCURACY	SPECIFICITY	SENSITIVITY
Naïve Bayesian Classifier	73.17%	60.00%	80.80%
SVM	73.20%	46.70%	88.50%
Logistic Regression	68.29%	33.30%	88.50%

IV. CONCLUSION AND DISCUSSION

The methods described here have improved quality, reduced labor time, and lowered overall cost for generation of protein profiles derived from a large sample set. Raspap has done this by adapting the SELDI protein chip processing, matrix application, and subsequent data analysis to an automated robotic and analysis components.

When using a protein chip bioprocessor for large-scale experiments, accurate manual pipetting is hampered due to the small apertures of the protein array wells (which can be less than 30 mm). To avoid contamination, the surface of the chip arrays cannot not be touched throughout the entire procedure. Also, variations in the application of matrix molecules have detrimental effects on the consistency of mass spectra. Therefore, in large-scale experiments, automation helps to circumvent these problems. This is also reflected in the significantly better CVs for the automated procedure compared to the manual one. The comparison of protein profiles from large-scale studies can also be facilitated via standardization and optimization of the process. Doing a frequent performance check of the spectrometer is also recommended for large scale studies.

With regard to the analysis side, the BAP component of the Raspap system has proven to yield fast and actionable results. By doing pipelined analyses and comparing different methods in an automated manner, the most striking patterns can emerge and be presented to researchers rapidly. BAP does this while interacting directly with the scientists and giving feedback to engineers.

Through use of the large cancer/control dataset, the Raspap method was employed successfully- from robotics to analysis. In this paper, predictive proteomic profiles results were discussed. However, other methods such as clustering and feature selection (i.e. specific protein peaks within the profile that are important in predicting the diagnosis) are also part of the pipeline. In fact, a number of such features were selected for biological protein identification in the case of this particular cancer dataset. Currently, collaborators on this project are in the process of publishing new biologic results based on these novel protein results which have been found to be clinically relevant.

V. References

[1] T. W. Hutchens and T. T. Yip, "New desorption strategies for the mass spectrometric analysis of macromolecules," *Rapid Communications in Mass Spectrometry*, vol. 7, pp. 576-580, 1993.

[2] E. F. Petricoin, 3rd, D. K. Ornstein, C. P. Paweletz, A. Ardekani, P. S. Hackett, B. A. Hitt, A. Velassco, C. Trucco, L. Wiegand, K. Wood, C. B. Simone, P. J. Levine, W. M. Linehan, M. R. Emmert-Buck, S. M. Steinberg, E. C. Kohn, and L. A. Liotta, "Serum proteomic patterns for detection of prostate cancer," *J Natl Cancer Inst*, vol. 94, pp. 1576-8, 2002.

[3] M. Woodford, "Automation of an experiment in inflammation: client/server laboratory automation for pharmaceutical R&D," presented at Proceedings of the 16th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. Part 2 (of 2), Nov 3-6 1994, Baltimore, MD, USA, 1994.

[4] W. Gecks and S. T. Pedersen, "Robotics--An efficient tool for laboratory automation," *IEEE Transactions on Industry Applications*, vol. 28, pp. 938-944, 1992.

[5] M. A. R. Meier, B.-J. de Gans, A. M. J. van der Berg, and U. S. Schubert, "Automated multiple-layer spotting for matrix-assisted laser desorption/ionization time-of-flight mass spectrometry of synthetic polymers utilizing ink-jet printing technology," *Rapid Communications in Mass Spectrometry*, vol. 17, pp. 2349-53, 2003.

[6] Raspap Internet Site, http://www.chip.org/proteomics/raspap/.

[7] I. Joliffe, *Principal Component Analysis*. New York, NY: Springer-Verlag, 1986.

[8] J. Hartigan and M. Wong, "Algorithm AS136: A k-means clustering algorithm.," *Applied Statistics*, vol. 28, pp. 100-108, 1979.

[9] T. M. Mitchell, *Machine Learning*. New York: McGraw-Hill, 1997.

[10] B. MK and L. PH, " PSA in the screening,staging and follow up of early-stage prostate cancer," *World J Urol*, vol. 7, pp. 7-11., 1989.