32nd Annual International Conference of the IEEE EMBS
Buenos Aires, Argentina, August 31 - September 4, 2010

SVM-Based Spectral Matching for Metabolite Identification

Bin Zhou, Amrita K. Cheema, and Habtom W. Ressom, *Senior Member, IEEE*

***Abstract*—Mass spectrometry-based metabolomics is getting mature and playing an ever important role in the systematic understanding of biological process in conjunction with other members of "-omics" family. However, the identification of metabolites in untargeted metabolomics profiling remains a challenge. In this paper, we propose a support vector machine (SVM)-based spectral matching algorithm to combine multiple similarity measures for accurate identification of metabolites. We compared the performance of this approach with several existing spectral matching algorithms on a spectral library we constructed. The results demonstrate that our proposed method is very promising in identifying metabolites in the face of data heterogeneity caused by different experimental parameters and platforms.**

I. Introduction

METABOLOMICS is the comprehensive and quantitative assessment of low molecular weight analytes (<1500Da) that define the metabolic status of an organism under a given condition [1]. In complementation with genomics, transcriptomics, and proteomics, the direct measurement of metabolite expression is essential in the systematic understanding of biological process. Metabolomics is increasingly enjoying widespread applications in areas such as functional genomics, identification of the onset and progression of disease, pharmacogeniomics, nutrigenomics, and systems biology [2-5].

Because of its sensitivity and coverage, mass spectrometry (MS) is a favorable technology for metabolomics study. Chromatography is often coupled to mass spectrometer to achieve further separation of the sample. Both gas chromatography (GC) and liquid chromatography (LC) have been used in metabolomics studies [2, 6].

One major bottleneck for current MS-based metabolomics is the identification of metabolites. In untargeted metabolomics, each detected compound is represented by a triplet of m/z, retention time and intensity. A common approach currently used for metabolite identification is to search the m/z value of detected peaks against a database (or databases). Several databases have been assembled during the past years [7-9]. The molecules in the database with a molecular weight within a specified tolerance to the query m/z value are retrieved as putative identifications of the compounds. However, there are some drawbacks for the mass-based searching method. First, it has been shown that even with an accuracy of 1ppm, which is a remarkably higher accuracy than most of the available platforms can achieve, it is still not sufficient for unambiguous metabolite identification [10]. Second, the isomers which have the same elemental composition but different structures have the same molecular weight. Thus, mass-based metabolite identification methods cannot discriminate isomers. To improve the identification of metabolites, additional information is needed such as retention time or fragment pattern. The latter is obtained by selecting a particular m/z from the first MS scan. The selected molecule is fragmented through collision induced dissociation. The resulting fragments are measured by the second MS scan. This approach provides us with a unique fingerprint for the compound. The fingerprint can be used for identification by comparing it with MS/MS spectra acquired from authentic compounds. Databases have begun to assemble the MS/MS spectra for authentic compounds using various platforms [9, 11].

To identify the correct metabolite from a large volume of MS/MS spectra, a proper comparison or scoring scheme is needed. The National Institute of Standards and Technology (NIST) has developed a scoring algorithm for compound identification by GC-MS [12]. It has also been modified for LC-MS-MS spectra matching by MassBank [13]. Another similar spectral matching algorithm was previously developed for peptide identification and integrated into the open-source software SpectraST [14]. While these algorithms perform well for spectra generated in highly-controlled environment, their performances degrade when spectra are generated from different labs using different platforms or with different parameters. Since it is costly and impractical to acquire the spectra under all possible conditions, robust spectral matching algorithms are needed for a general spectral library to be useful.

In this paper, we collected MS/MS spectra for 21 metabolites from both our in-house data and publicly available data from the Human Metabolite Database (HMDB). We utilized a support vector machine (SVM) to incorporate both peak and profile similarity measures for spectral matching. We compared the identification performance of our proposed approach with other algorithms (NIST, MassBank, and SpectraST) and the correlation method. We observed that the proposed approach can achieve 7% to 10% improvement on identification performance.

Bin Zhou is with the Department of Electrical and Computer Engineering at Virginia Polytechnic Institute and State University, Falls Church, VA 22043 USA (e-mail: zhoubin@vt.edu).

Amrita K. Cheema is with the Department of Oncology, Georgetown University, DC, 20057 USA (e-mail: akc27@georgetown.edu).

Habtom W. Ressom is with the Department of Oncology, Georgetown University, DC, 20057 USA (corresponding author, phone: 202-687-2283; fax: 202-687-0227; e-mail: hwr@georgetown.edu).

II. SPECTRA MATCHING ALGORITHMS

During the past decade, several spectral matching algorithms have been developed for various applications and platforms [13-16]. While there are multiple forms of scoring algorithms for spectral matching, they are primarily some variations of dot product. The dot product of a query spectrum and a library spectrum intrinsically measures the correlation between the two spectra. The library spectrum which has the highest correlation with the query spectrum is considered to be the right identification.

While correlation generally performs well when the spectra are from highly controlled experiment, it will degrade remarkably in real situation when the spectra are from different sources. The underlying reason is that different platforms have different analyzing and detection mechanisms which cause the intensity of fragments in MS/MS spectra to vary. Also experimental parameters such as collision energy have a significant impact on the intensity profile of a fragment spectrum. An example is shown in Fig. 1, where two spectra of the same metabolite exhibit different spectral profiles when acquired under different conditions.

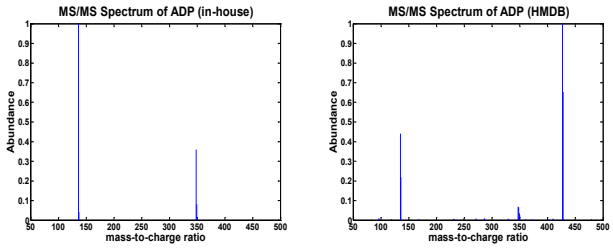

Fig. 1. The MS/MS spectra for metabolite ADP from the in-house data (left panel) and from HMDB database (right panel).

As a result, when there is a high heterogeneity of data, correlation or dot product alone is not sufficient to measure the similarity between two spectra. Other measures of similarity are needed. From the example given in Fig. 1, we observe that while the profiles of the two spectra are not similar, they share peaks appearing at the same positions over the m/z range. To take advantage of this observation, we propose to utilize "peak similarity", which measures the impact of the common peaks on the spectral comparison. Specifically, we define the following two measures of peak similarity:

$$N_{match} = N_{U\&L} / \min(N_U, N_L) \quad 2.1$$

$$E_{match} = \sum_{i=1}^{N_{U\&L}} (A_{L,i} A_{U,i}) \quad 2.2$$

where N_{match} is the normalized number of common peaks between the two spectra and E_{match} is the total energy of common peaks. A_L and A_U are the relative intensity of peak lists of the library spectrum and query spectrum. N_U and N_L are the length of the peak list of query spectrum and library spectrum respectively. $N_{U\&L}$ is the length of the common peaks appearing in both spectra.

We propose to combine the above two peak similarity measures with a measure of profile similarity. To measure the profile similarity of two spectra, we use the Pearson correlation coefficient between the spectra. The correlation coefficient must be calculated between the vectors of the same length. Since the lengths of two spectra (peak list) are rarely equal, the spectra must be re-sampled before the calculation. We use the peak preserving re-sampling to re-sample the spectra [17]. The signal is re-constructed using a Gaussian kernel and the intensity at an m/z value is the maximum intensity of any contributing peaks. The Pearson correlation is then calculated using the re-sampled spectra.

The final identification is based on the overall consideration of both profile and peak similarity measures. We formulate the identification problem as a classification problem. Each comparison between two spectra is a sample for classification while the similarity measures are the features of the sample. And the label is binary: the sample is positive for comparison between the same metabolites while negative for comparison between different metabolites. The proposed metabolite identification algorithm is illustrated in Fig. 2. Pair-wise comparisons are performed between query spectrum and the spectrum from the library. The normalized number of common peaks, the total energy of common peaks, and the Pearson correlation are utilized to train a SVM classifier to decide if the two spectra represent the same metabolite.

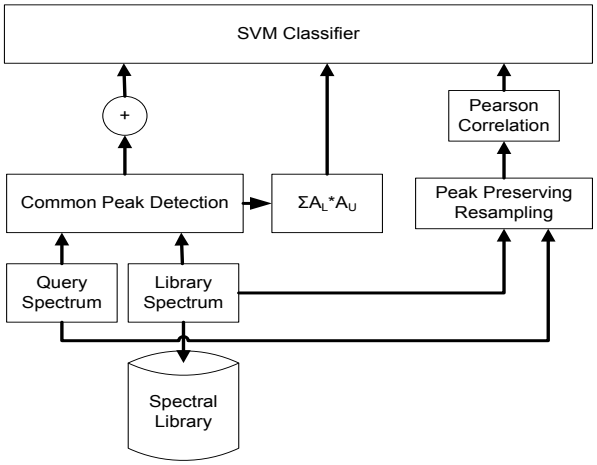

Fig. 2. The algorithm diagram for the SVM-based approach.

In [18], a similar scheme was used for peptide identification, however, the author used linear discriminant function to combine the multiple similarity measures. By examining the data we have, we think the separating plane for positive and negative samples should be non-linear. The reason is that different similarity measures have different dynamic ranges and there is no explicit way to normalize them to make the data linearly separable. SVM with radial basis kernel function is a convenient and popular way to solve this kind of non-linear separation problems [19].

The metabolite identification system is intrinsically an information retrieval system. One common characteristics of information retrieval system is the imbalance of the samples. In the study, more negative samples are present than positive

samples. It is well-known that such imbalance will have an adversary impact on the performance of SVM classifier [20]. Several ways have been devised to correct for sample imbalance, including re-sampling (either down-sampling or over-sampling), cost-sensitive learning or ensemble learning [20-22]. Here, we use a random down-sampling approach on the majority group to acquire a balanced dataset to train the SVM classifier.

III. Experiment Results

A. Data acquisition

The in-house data consisted of 21 metabolites with molecular weights ranging from 107 to 428 Da. For each metabolite, authentic compound was purchased to acquire the MS/MS spectra using an ultra performance liquid chromatography quadrupole time of flight (UPLC-QTOF, Waters) instrument. The collision energy was tuned for each individual compound to acquire MS/MS spectra with a reasonable number of fragments. For some of the metabolites, more than one spectrum was acquired. Totally, our in-house data comprised of 45 MS/MS spectra representing 21 metabolites. The MS/MS spectra for the same 21 metabolites were also retrieved from the HMDB database. The spectra were acquired using a high performance liquid chromatography triple quadrupole (HPLC-QqQ, Waters) instrument with a collision energy of 10eV. For each metabolite, only one MS/MS spectrum was available in HMDB.

B. Experiment design

To evaluate the performance of different algorithms, a 3-fold cross validation was performed. The 21 metabolites were randomly divided into two groups. In the training set, we had spectra for 14 metabolites from both the in-house data and the HMDB database. In the testing set, we had spectra for 7 metabolites from the in-house data and the HMDB database. The purpose of this stratification is to make sure that we have some degree of data heterogeneity in both training and testing sets. The training set was used to train a model to discriminate positive samples and negative samples. For a typical scoring method that uses only one variable to measure spectral similarity, the model is a threshold for the score, which maximizes the F-measure in the training set. For the proposed method, the model is a SVM classifier trained on the training set. The trained model was applied to the testing set to evaluate the identification performance. This procedure was repeated 100 times to measure the performance of algorithms. In addition to the four algorithms previously introduced, we used the Pearson correlation coefficient as a score for performance comparison.

Because we were particularly interested in evaluating the performance of various metabolites identification algorithms on heterogeneous datasets, we carried out two experiments. In Experiment I, we conducted identification using data from different sources. Specifically, the query spectra from the in-house data were searched against a library composed of spectra from HMDB, or vice versa. In Experiment II, spectra from different sources were mixed together to form a mixed dataset and a spectrum in the dataset was searched against other spectra in the mixed dataset.

Because the data are highly imbalanced with much more negative samples, accuracy only is not enough to measure the identification performance. Thus we utilized F-measure notion from information retrieval context to measure the performances of the algorithms, which is defined as

$$F = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}, \quad \text{Precision} = \frac{TP}{TP+FP}, \quad \text{Recall} = \frac{TP}{TP+FN}$$

where TP, FP, FN are the number of true positive, false positive, and false negative, respectively. Since we have more negative samples than positive samples and we are mainly concerned about the correct identification of positive samples. Therefore, in our study, F-measure is more suitable for performance evaluation than accuracy.

C. Experiment results

The Pearson correlation coefficients for positive and negative samples in Experiments I and II are shown in Figs. 3 and 4, respectively. These figures illustrate the necessity to induce more similarity measures in addition to correlation. The comparisons between different metabolites (negative samples) generally show very small correlation coefficients as expected. However, the comparisons between the same metabolites (positive samples) span a large range of correlation coefficients. For some of them, the spectral profiles from different platforms and experiments are similar, while for others there is a large variation between the spectra.

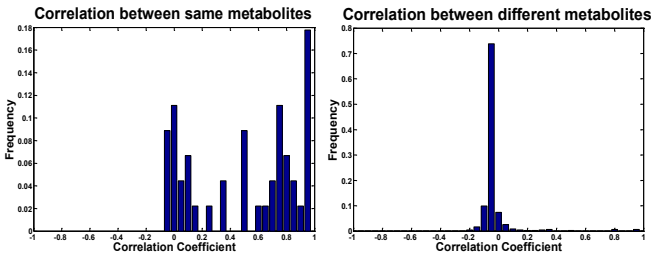

Fig. 3. Correlation coefficients for comparison between same metabolites (left panel) and different metabolites (right panel) in Experiment I.

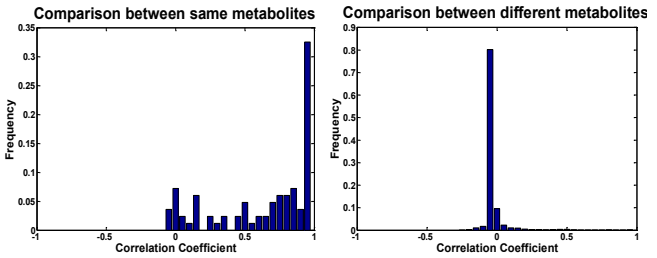

Fig. 4. Correlation coefficients for comparison between same metabolites (left panel) and different metabolites (right panel) in Experiment II.

Table I presents the accuracy and F-measure of five spectral matching algorithms. Among the five algorithms, SVM gives the best performance on both accuracy and F-measure. While the accuracies of other algorithms are

comparable, SVM achieved about 7% to 10% increase on F-measure. SVM also outperforms the other algorithms in terms of the area under the curve (AUC) of the precision-recall graph as shown Table I, Fig. 5, and Fig. 6.

Table I. The performance of spectral matching algorithms

Experiment	Method	F-measure (%)	Accuracy (%)	AUC (%)
I	NIST	74.0	93.8	79.7
	MassBank	72.1	93.3	83.5
	SpectraST	69.3	92.1	72.0
	Correlation	73.2	93.2	75.1
	SVM	**80.7**	**94.6**	**86.9**
II	NIST	77.7	95.3	84.3
	MassBank	77.3	95.2	86.1
	SpectraST	75.7	94.6	79.7
	Correlation	76.7	95.1	87.1
	SVM	**85.1**	**96.3**	**90.1**

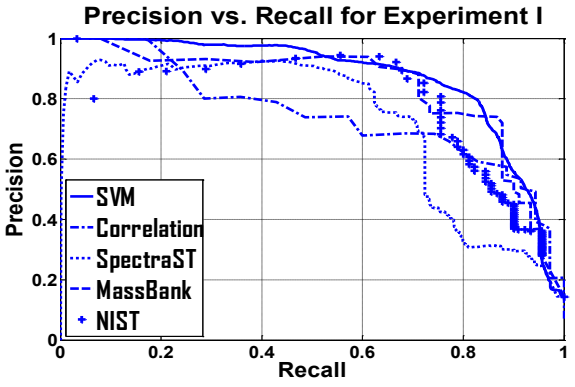

Fig. 5. Precision-recall graph for the spectral matching algorithms in Experiment I.

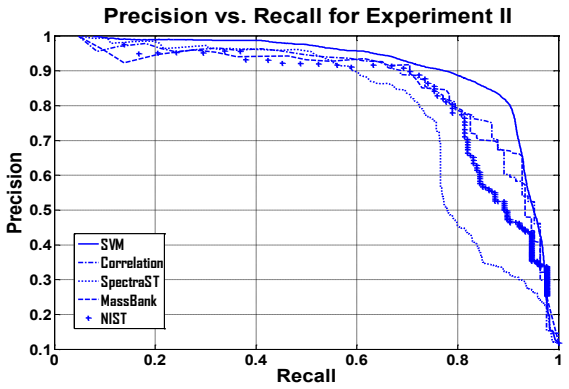

Fig. 6. Precision-recall graph for the spectral matching algorithms in Experiment II.

IV. Conclusion

In this paper, we propose two metrics that measure the similarity of peaks present in two MS/MS spectra. The two metrics are combined with correlation through SVM. We demonstrate the ability of this approach to give more accurate identification of metabolites by comparing it with several other spectral matching algorithms. We observe that the dot product alone is not sufficient for identification in heterogeneous data. Our results indicate that the proposed approach is very promising and outperforms the existing spectral matching algorithms for metabolite identification.

V. Acknowledgement

This work was supported by the National Science Foundation Grant (IIS-0812246).

References

[1] R. Goodacre, *et al.*, "Metabolomics by numbers: acquiring and understanding global metabolite data," *Trends Biotechnol,* vol. 22, pp. 245 - 252, 2004.

[2] C. Chen, *et al.*, "LC-MS-Based Metabolomics in Drug Metabolism," *Drug Metabolism Reviews,* vol. 39, pp. 581-597, 2007.

[3] A. Sreekumar, *et al.*, "Metabolomic profiles delineate potential role for sarcosine in prostate cancer progression," *Nature,* vol. 457, pp. 910-914, 2009.

[4] P. Yin, *et al.*, "Metabonomics Study of Intestinal Fistulas Based on Ultraperformance Liquid Chromatography Coupled with Q-TOF Mass Spectrometry (UPLC/Q-TOF MS)," *Journal of Proteome Research,* vol. 5, pp. 2135-2143, 2006.

[5] A. D. Patterson, *et al.*, "UPLC-ESI-TOFMS-Based Metabolomics and Gene Expression Dynamics Inspector Self-Organizing Metabolomic Maps as Tools for Understanding the Cellular Response to Ionizing Radiation," *Analytical Chemistry,* vol. 80, pp. 665-674, 2008.

[6] P. Jonsson, *et al.*, "A Strategy for Identifying Differences in Large Series of Metabolomic Samples Analyzed by GC/MS," *Analytical Chemistry,* vol. 76, pp. 1738-1745, 2004.

[7] D. S. Wishart, *et al.*, "HMDB: the Human Metabolome Database," *Nucleic acids research,* vol. 35, pp. D521 - 6, 2007.

[8] Q. Cui, *et al.*, "Metabolite identification via the Madison Metabolomics Consortium Database," *Nat Biotech,* vol. 26, pp. 162-164, 2008.

[9] C. A. Smith, *et al.*, "METLIN: A Metabolite Mass Spectral Database," *Therapeutic Drug Monitoring,* vol. 27, pp. 747-751, 2005.

[10] T. Kind and O. Fiehn, "Metabolomics database annotations via query of elemental compositions: Mass accuracy is insufficient even at less than 1 ppm," *BMC Bioinformatics,* vol. 7, p. 234, 2006.

[11] D. S. Wishart, *et al.*, "HMDB: a knowledgebase for the human metabolome," *Nucl. Acids Res.,* vol. 37, pp. D603-610, January 1, 2009 2009.

[12] N. Schauer, *et al.*, "GC-MS libraries for the rapid identification of metabolites in complex biological samples," *FEBS Letters,* vol. 579, pp. 1332-1337, 2005.

[13] H. Hisayuki, "Comparison of ESI-MS Spectra in MassBank Database," 2008, pp. 853-857.

[14] L. Henry, *et al.*, "Development and validation of a spectral library searching method for peptide identification from MS/MS," *PROTEOMICS,* vol. 7, pp. 655-667, 2007.

[15] S. E. Stein, "Estimating probabilities of correct identification from results of mass spectral library searches," *Journal of the American Society for Mass Spectrometry,* vol. 5, pp. 316-323, 1994.

[16] S. E. Stein and D. R. Scott, "Optimization and testing of mass spectral library search algorithms for compound identification," *Journal of the American Society for Mass Spectrometry,* vol. 5, pp. 859-866, 1994.

[17] L. Lars, "Visual Analysis of Gel-Free Proteome Data," *IEEE Transactions on Visualization and Computer Graphics,* vol. 12, pp. 497-508, 2006.

[18] J. P. Dworzanski, *et al.*, "Identification of Bacteria Using Tandem Mass Spectrometry Combined with a Proteome Database and Statistical Scoring," *Analytical Chemistry,* vol. 76, pp. 2355-2366, 2004.

[19] T. Hastie, *et al.*, *The Elements of Statistical Learning*: Springer, 2001.

[20] D. R. Musicant, *et al.*, "Optimizing F-Measure with Support Vector Machines," in *Proceedings of the Sixteenth International Florida Artificial Intelligence Research Conference*, 2003, pp. 356-360.

[21] G. H. Nguyen, *et al.*, "Learning Pattern Classification Tasks with Imbalanced Data Sets, Pattern Recognition," in *Pattern Recognition*, P.-Y. Yin, Ed., ed: INTECH, 2009, pp. 193-208.

[22] Y. Tang, *et al.*, "SVMs modeling for highly imbalanced classification," *Trans. Sys. Man Cyber. Part B,* vol. 39, pp. 281-288, 2009.