Published in final edited form as:

Conf Proc IEEE Eng Med Biol Soc. 2015 August ; 2015: 8185–8188. doi:10.1109/EMBC.2015.7320294.

Integrative Analysis of LC-MS based Glycomic and Proteomic Data

Minkun Wang,

Department of Electrical and Computer Engineering, Virginia Tech, Arlington, VA 22203, USA

Guoqiang Yu, and

Department of Electrical and Computer Engineering, Virginia Tech, Arlington, VA 22203, USA

Habtom W. Ressom, IEEE* [Senior Member]

Department of Electrical and Computer Engineering, Virginia Tech, Arlington, VA 22203, USA

Minkun Wang: wmkevin@vt.edu; Guoqiang Yu: yug@vt.edu

Abstract

Studies associating changes in the levels of glycans and proteins with the onset of cancer have been widely investigated to identify clinically relevant diagnostic biomarkers. Advances in liquid chromatography mass spectrometry (LC-MS) have enabled high-throughput identification and quantitative analysis of these biomolecules. While results from separate analyses of glycans and proteins have been reported widely, the mutual information obtained by combining the two has been relatively unexplored. In this study, we investigate integrative analysis of glycans and proteins to take advantage complementary information to improve the ability to distinguish cancer cases from controls. Specifically, SVM-RFE algorithm is utilized to select a panel of N-glycans and proteins from LC-MS data previously acquired by analysis of sera from two cohorts in a liver cancer study. Improved performances are observed by integrative analysis compared to separate glycomic and proteomic studies in distinguishing liver cancer cases from patients with liver cirrhosis.

I. Introduction

Glycosylation is one of the most common post-translational modifications of proteins. Altered patterns of glycosylation have been associated with various diseases and many currently used cancer biomarkers. In particular protein glycosylation is relevant to liver pathology because of the major influence of this organ on the homeostasis of blood glycoproteins. Characterizing glycan modifications of proteins in complex proteomes is challenging as glycosylation can occur on multiple sites of peptides involving the attachment of different glycans to each site. An alternative strategy to the analysis of glycoproteins is the study of proteins and protein-associated glycans [1, 2].

We previously performed separate analyses of proteins and N-glycans released from proteins in sera by using liquid chromatography coupled with mass spectrometry (LC-MS) [3]. We

*Corresponding author: Habtom W. Ressom (hwr@georgetown.edu).

detected N-glycans and proteins significantly altered in hepatocellular carcinoma (HCC) cases compared to patients with liver cirrhosis using univariate statistical methods. However, multivariate statistical or machine learning methods are desirable to improve the ability to discriminate the cases from controls by taking advantage of the mutual information within the N-glycans and proteins themselves as well as the combination of the two. The analysis will allow investigating if the synergy of the two omic studies leads to improved performance in distinguishing cases from controls compared to the a single omic study.

In this paper, we investigate two datasets we previously generated by LC-MS based serum glycomic and proteomic studies to identify N-glycans and proteins significantly altered in HCC versus patients with liver cirrhosis. The goal of the investigation in this paper is, however, to evaluate the improvement achieved in disease classification by integrating the two datasets using machine learning methods. Specifically, support vector machine-recursive feature elimination (SVM-RFE) is used to select an optimal set of features that leads to highly discriminant classifier [4]. This not only helps recognize relevant patterns in the feature space, but also reduces dimensionality to overcome the risk of overfitting. We apply this SVM-RFE approach to proteomic and glycomic data from a liver cancer study. Through a 10-fold cross validation, we evaluated the classification performances of the features selected from each omic studies as well as the combined features. We observed that improved performances can be achieved through the integrative analysis compared to separate glycomic and proteomic studies.

The remaining part of this paper is organized as follows. Section II briefly summarizes the experimental design used for acquisition of glycomic and proteomic datasets. Also, this section describes our normalization, feature Selection, and disease classification methods used for integrative analysis of the two datasets. Section III presents the results we obtained in selecting optimal features from each dataset as well as the integrated glycomic and proteomic dataset. Section IV concludes the paper with summary and future goals.

II. Materials and Methods

A. Experimental Design

The proposed integrative analysis is performed on glycomic and proteomic datasets we previously acquired by LC-MS based analysis of serum samples from HCC cases and patients with liver cirrhosis recruited in Egypt and the U.S [5, 6]. The participants in Egypt and the U.S. were recruited through protocols approved by the Ethics Committee at Tanta University Hospital and the Institutional Review Board at Georgetown University respectively. Specifically, adult patients were recruited from the outpatient clinics and inpatient wards of the Tanta University Hospital (TU cohort) in Tanta, Egypt and from the hepatology clinics at MedStar Georgetown University Hospital (GU cohort) in Washington, DC, USA. The TU cohort consists of a total of 89 subjects (40 HCC cases and 49 patients with liver cirrhosis), and the GU cohort comprises of 116 subjects (57 HCC cases and 59 patients with liver cirrhosis).

Fig. 1 depicts the overall workflow of our experimental design. Briefly, N-glycans and proteins in human sera were analyzed by using LC-MS based untargeted and targeted

methods. Significantly altered N-glycans and proteins were selected from the untargeted analysis and further confirmed through a targeted analysis by multiple reaction monitoring (MRM). Two datasets were obtained from the glycomic and proteomic studies following analysis of the LC-MS data by GPA [7] and Skyline [8], respectively. Results from univariate statistical analysis have been previously reported in [5, 6]. In the following, we introduce how we normalize and integrate the two datasets for feature selection and disease classification.

B. Normalization, Feature Selection, and Disease Classification

Observing that the distributions of measurements from LC-MS data were skewed (Fig. 2a), we log-transformed the two data sets to resemble normal distributions (Fig. 2b). To make the two matrices compatible for integration, we performed Z-score normalization before integrating the two data sets (Fig. 2c). This step ensures features from protein and glycan lists are treated equally in the feature selection procedure.

Linear SVMs were trained to classify samples in case and control groups using features from the proteomic and glycomic studies separately as well as combining features from both studies. Equation (1) presents the decision function in SVM model for an input sample $\mathbf{x}_t$

$$D(\mathbf{x}_t) = \mathbf{w} \cdot \mathbf{x}_t + b, \text{ where } \mathbf{w} = \sum_k \alpha_k y_k \mathbf{x}_k \text{ and } b = \langle y_k - \mathbf{w} \cdot \mathbf{x}_k \rangle. \quad (1)$$

The feature weight vector $\mathbf{w}$ determined by support vectors is employed as feature ranking criterion by the recursive feature elimination (RFE) algorithm [4]. SVM-RFE eliminates redundant features iteratively and yields better and more compact feature subsets. The major steps include 1) training the SVM classifier; 2) ranking the features according to weight vector $\mathbf{w}$ of the learned SVM; 3) eliminating features with the smallest ranking criterion; 4) retraining SVM model with the remaining features; 5) estimating the performance of the model using cross-validation to check if the optimal subset is obtained. In this paper, we applied SVM-RFE to select highly discriminative sets of N-glycans and proteins, as well as a combination of the two in a panel.

III. Results and Discussion

Targeted analyses of 82 N-glycans (characterized by the number five monosaccharides: GlcNAc, mannose, galactose, fucose, and NeuNAc.) and 101 proteins (represented by Uniprot IDs) were performed using a Dionex 3000 Ultimate nano-LC system (Dionex Sunnyvale, CA) interfaced to TSQ Vantage mass spectrometer (Thermo Scientific, San Jose CA). In addition to the two separate datasets obtained from targeted analysis of N-glycans and proteins, a third dataset obtained by concatenating the two datasets was analyzed by SVM-RFE following log-transformation and Z-score normalization (Fig. 2).

SVM-RFE models were trained for these three datasets respectively. We started from the whole feature list in each dataset, and eliminated one feature in each iteration step till feature set was empty. 10-fold cross validation was employed to evaluate the average classification performance (i.e., accuracy, sensitivity, and specificity) at each step. Figures 3a and 3b depict the classification accuracy achieved at each iteration step for the top 50 features

selected from the three datasets in the TU and GU cohorts, respectively. Also, the figures show the optimal number of features that leads to the best classification accuracy. We observed that, in most iteration steps, features selected from the integrated dataset yield higher accuracies compared to the same number of features selected from either the glycomic or proteomic dataset.

Receiver operating characteristics (ROC) curves were estimated by varying the SVM threshold parameter ($y_f = \hat{w} \cdot x - b$). The 95% confidence intervals of area under the ROC (AUC) were calculated using bootstrap method with 1000 resampled replicates.

Table I shows the results for the best classification performance with optimal subset of features in each dataset of the TU cohort. As shown in the table, SVM-RFE selected 29 out of 82 N-glycans and 15 out of 101 proteins as the optimal number of features. Among these, 13 glycans and 5 proteins were also selected as significantly altered in cases versus controls through univariate statistical test [5, 6]. Out of 183 integrated features, 7 proteins and 2 N-glycans in a panel were selected by SVM-RFE. The panel includes 2 that were also found significant in the univariate statistical analysis. The integrative analysis led to a significantly smaller number of features with a slight improvement on the disease classification accuracy compared to those selected by analysis of individual datasets. This phenomenon is observed consistently across the entire iteration steps, as illustrated in Fig. 3a. Similar results are obtained in the GU cohort (Table II), in which SVM-RFE selected 18 proteins and 5 N-glycans in a panel yielded better performance than 22 proteins or the 8 glycans selected by analysis of individual datasets. Among the 23 features selected by the integrative analysis, four N-glycans and 10 proteins were also reported as significant by univariate statistical analysis. As shown in Fig. 3b, the integrative analysis yielded improved performance compared to the analysis based on the individual datasets in the majority of the iteration steps. In both cohorts, we captured features with synergic contributions to the discrimination, which provide complementary information to univariate analysis. Although we did not observe overlapping features between the optimal sets of features in the two cohorts, we were able to achieve AUCs greater than 0.73 when we trained SVMs based on the data the integrated panel learned from TU cohort and tested it on the GU cohort, and vice versa.

Finally, we investigated the performance for each dataset by setting the feature size to five. We compared the performances of the best five features selected by SVM-RFE from each of the three datasets (Table III). While the integrative analysis outperformed the analysis based on individual dataset in TU cohort, both the integrated features and the protein features led to similar performances in the GU cohort.

IV. Conclusion and Future Work

In this study, we investigated the benefit of an integrative analysis of glycomic and proteomic datasets in improving our ability to distinguish HCC cases from patients with liver cirrhosis. Through SVM-RFE, a panel of features was selected from 82 N-glycans and 101 proteins previously quantitated by LC-MS based targeted analysis. Complementary to univariate statistical methods, the integrative analysis utilizes mutual information among

features to select a panel of features with improved ability to discriminate the two biological groups. In this study, we observe that features selected by merging the glycomic and proteomic datasets lead to better disease classification accuracy compared to those selected from one of the two datasets. We would like to emphasize that the improvement achieved by the integrative analysis was observed not only in using SVM-RFE, but also through other methods such as a sequential feature selection coupled with quadratic discriminant analysis. We believe that integrative analysis by multivariate analysis combined with pathway-centric and network-based analyses of multi-omic data will help not only features that improved disease classification but also biomolecules that give insight into the molecular mechanisms of complex diseases.

Acknowledgments

This work is in part supported by the National Institutes of Health Grants R01CA143420 and R01GM086746 awarded to HWR.

References

1. Fuster MM, Esko JD. The sweet and sour of cancer: Glycans as novel therapeutic targets. *Nat. Rev. Cancer*. 2005; 5(7):526–542. [PubMed: 16069816]
2. Blomme B, Van Steenkiste C, Callewaert N, Van Vlierberghe H. Alteration of protein glycosylation in liver diseases. *J Hepatol*. 2009; 50(3):592–603. [PubMed: 19157620]
3. Kulasingam V, Diamandis EP. Strategies for discovering novel cancer biomarkers through utilization of emerging technologies. *Nat. Clin. Pract. Oncol*. 2008; 5(10):588–599. [PubMed: 18695711]
4. Guyon I, Weston J, Barnhill S, Vapnik V. Gene selection for cancer classification using support vector machines. *Mach. Learning*. 2002; 46(1–3):389–422.
5. Tsai T, Wang M, Di Poto C, Hu Y, Zhou S, Zhao Y, Varghese RS, Luo Y, Tadesse MG, Ziada DH. LC–MS profiling of N-glycans derived from human serum samples for biomarker discovery in hepatocellular carcinoma. *Journal of Proteome Research*. 2014; 13(11):4859–4868. [PubMed: 25077556]
6. Tsai T, Song E, Zhu R, Di Poto C, Wang M, Luo Y, Varghese RS, Tadesse MG, Ziada DH, Desai CS, Shetty K, Mechref Y, Ransom HW. LC–MS/MS based Serum Proteomics for Identification of Candidate Biomarkers for Hepatocellular Carcinoma. *Proteomics*. 2015 In Press.
7. Wang, M.; Yu, G.; Mechref, Y.; Ransom W, H. GPA: An algorithm for LC/MS based glycan profile annotation. *IEEE International Conference on Bioinformatics and Biomedicine Workshop (BIBM 2013)*; Shanghai, China. 2013 Dec.
8. MacLean B, Tomazela DM, Shulman N, Chambers M, Finney GL, Frewen B, Kern R, Tabb DL, Liebler DC, MacCoss MJ. Skyline: An open source document editor for creating and analyzing targeted proteomics experiments. *Bioinformatics*. 2010; 26(7):966–968. [PubMed: 20147306]

Figure 1.
Workflow of integrative analysis of N-glycans and proteins.

Figure 2.

The distributions of raw glycomic (orange) and proteomic (cyan) datasets (a); log-transformed data (b); data after log-transformation and Z-score normalization.

Figure 3.

Classification accuracy at each iteration step for the top 50 features from glycomic (green), proteomic (blue), and integrated datasets (red) in the TU and GU cohorts. The optimal numbers of features (indicated by triangles) correspond to the best classification accuracy (indicated by circles).

TABLE I

Performance Comparison Based on the Optimal Number of Features Selected in the TU Cohort

TU Cohort	<i>Glycomic</i>	<i>Proteomic</i>	<i>Integrated</i>
Accuracy	0.77	0.84	0.87
Sensitivity	0.82	0.83	0.90
Specificity	0.75	0.84	0.85
AUC (95% CI)	0.87 (0.78, 0.93)	0.93 (0.83, 0.97)	0.92 (0.81, 0.97)
Optimal Number of Features	29/82	15/101	9/183
Selected Features^b	[25000]	[43000] ^a	P01024 ^a
	[53111] ^a	[34100] ^a	P02743
	[63402] ^a	[43202] ^{a,c}	P02750
	[53313]	[53000] ^a	P02753 ^a
	[53323]	[33101]	P02763
	[34110]	[63403] ^a	P03952
	[53311] ^c	[53311] ^c	P04004 ^a
	[43110] ^a	[53010]	P05160
	[63413] ^a	[53302] ^a	P06727
	[53411]	[34101]	P0C0L4
	[63423]	[63404] ^a	P13598 ^a
	[53312]	[29000]	P13796
	[53101] ^a	[73514]	P22891 ^a
	[53201]	[43202] ^{a,c}	P27918
	[2 10 000]		P35858

^aSignificant (p value = 0.05) in univariate statistical analysis^bN-glycans are characterized by GlcNAc, mannose, galactose, fucose, and NeuNAc, and proteins are indicated by Uniprot IDs^cIsomers with different retention times.

TABLE II

Performance Comparison Based on the Optimal Number of Features Selected in the GU Cohort

GU Cohort	Glycomic	Proteomic	Integrated
Accuracy	0.77	0.88	0.91
Sensitivity	0.79	0.86	0.89
Specificity	0.75	0.91	0.93
AUC (95% CI)	0.83 (0.71, 0.91)	0.95 (0.89, 0.98)	0.96 (0.89, 0.99)
Optimal Number of Features	8/82	22/101	23/183
Selected Features^b		O75015 O75636 ^a	O75015 O75636^a
		P00748 ^a P01023 ^a	P01023^a P01034^a
	[43100] ^a	P01877 ^a P02741	P01877^a P02771^a
	[53313]	P02766 P02771 ^a	P04278 P05155
	[53000] ^a	P02790 P04278	P05452^a P08294
	[43212]	P05155 P05452	P13796 P41222^a
	[53411]	P06727 P13796	P61626^a Q13201^a
	[53312]	P27169 ^a P41222 ^a	Q15848^a Q96KN2
	[53200]	P49747 ^a P61626 ^a	[43100]^a [53313]
	[63434]	P61769 ^a Q15848 ^a	[53000]^a [43200]^a
		Q96KN2	[53411] [53200]
		Q9Y6R7 ^a	[53111]^a

^aSignificant (p value = 0.05) in univariate statistical analysis^bN-glycans are characterized by GlcNAc, mannose, galactose, fucose, and NeuNAc, and proteins are indicated by Uniprot IDs

TABLE III

Performance Comparison on the Top Ranking five Featured Selected in the TU Cohort and GU Cohorts

TU cohort	<i>Glycomic</i>	<i>Proteomic</i>	<i>Integrated</i>
Accuracy	0.68	0.79	0.83
Sensitivity	0.71	0.79	0.82
Specificity	0.67	0.79	0.85
AUC (95% CI)	0.77 (0.65, 0.59)	0.88 (0.77, 0.94)	0.89 (0.80, 0.95)
Number of Selected Features	5/82	5/101	5/183
Selected Features		<ul style="list-style-type: none"> Complement C3^a Serum amyloid P-component Alpha-1-acid glycoprotein 1 Coagulation factor XIII B chain Apolipoprotein A-IV 	<ul style="list-style-type: none"> Serum amyloid P-component Alpha-1-acid glycoprotein 1 Coagulation factor XIII B chain Apolipoprotein A-IV
GU cohort	<i>Glycomic</i>	<i>Proteomic</i>	<i>Integrated</i>
Accuracy	0.74	0.80	0.80
Sensitivity	0.74	0.82	0.82
Specificity	0.75	0.79	0.79
AUC (95% CI)	0.82 (0.70, 0.89)	0.85 (0.74, 0.92)	0.87 (0.77, 0.93)
Number of Selected Features	5/82	5/101	5/183
Selected Features		<ul style="list-style-type: none"> Ficolin-3^a Alpha-2-macroglobulin^a Plasma protease C1 inhibitor Tetranectin^a Prostaglandin-H 2 D-isomerase^a 	<ul style="list-style-type: none"> Ficolin-3^a Alpha-2-macroglobulin^a Plasma protease C1 inhibitor Sex hormone binding globulin
Annotation	N-acetylglucosamine , mannose , galactose , fucose , NeuNAc

Significant (p value = 0.05) proteins in univariate statistical analysis. N-glycans that found significant (p value = 0.05) in univariate statistical analysis are shown in boxes.