Journal of Biomedicine and Biotechnology • 2005:2 (2005) 80–86 • DOI: 10.1155/JBB.2005.80

RESEARCH ARTICLE

Functional Clustering Algorithm for High-Dimensional Proteomics Data

Halima Bensmail,[1] Buddana Aruna,[1] O. John Semmes,[2] and Abdelali Haoudi[2]

[1]*Department of Statistic Operation and Management Sciences (SOMS), The University of Tennessee, Knoxville, TN 37996, USA*
[2]*Department of Microbiology and Molecular Cell Biology, Eastern Virginia Medical School, Norfolk, VA 23507, USA*

Received 9 September 2004; revised 10 February 2005; accepted 14 February 2005

Clustering proteomics data is a challenging problem for any traditional clustering algorithm. Usually, the number of samples is largely smaller than the number of protein peaks. The use of a clustering algorithm which does not take into consideration the number of features of variables (here the number of peaks) is needed. An innovative hierarchical clustering algorithm may be a good approach. We propose here a new dissimilarity measure for the hierarchical clustering combined with a functional data analysis. We present a specific application of functional data analysis (FDA) to a high-throughput proteomics study. The high performance of the proposed algorithm is compared to two popular dissimilarity measures in the clustering of normal and human T-cell leukemia virus type 1 (HTLV-1)-infected patients samples.

INTRODUCTION

A variety of mass spectrometry-based platforms are currently available for providing information on both protein patterns and protein identity [1, 2]. Specifically, the first widely used such mass spectrometric technique is known as surface-enhanced laser desorption ionization (SELDI) coupled with time-of-flight (TOF) mass spectrometric detection [3, 4, 5]. The SELDI approach is based on the use of an energy-absorbing matrix such as sinapinic acid (SPH), large molecules such as peptides ionize instead of decomposing when subjected to a nitrogen UV laser. Thus, partially purified serum is crystallized with an SPH matrix and placed on a metal slide. Depending upon the range of masses the investigator wishes to study, there are a variety of possible slide surfaces; for example, the strong anion exchange (SAX) or the weak cation exchange (WCX) surface. The peptides are ionized by the pulsed laser beam and then traverse a magnetic-field-containing column. Masses are separated according to their TOFs as the latter are proportional to the square of the mass-to-charge (m/z) ratio. Since nearly all of the resulting ions have unit charge, the mass-to-charge ratio is in most cases a mass. The spectrum (intensity level as a function of mass) is recorded, so the resulting data obtained on each serum sample are a series of intensity levels at each mass value on a common grid of masses (peaks).

Proteomic profiling is a new approach to clinical diagnosis, and many computational challenges still exist. Not only are the platforms themselves still improving, but the methods used to interpret the high-dimensional data are developing as well [6, 7].

A variety of clustering approaches has been applied to high-dimensional genomics and proteomics data [8, 9, 10, 11]. Hierarchical clustering methods give rise to nested partitions, meaning the intersection of a set in the partition at one level of the hierarchy with a set of the partition at a higher level of the hierarchy will always be equal to the set from the lower level or the empty set. The hierarchy can thus be graphically represented by a tree.

Functional data analysis (FDA) is a statistical data analysis represented by smooth curves or continuous functions $\mu_i(t)$, $i = 1, \ldots, n$, where n is the number of observations and t might or might not necessarily denote time but might have a general meaning. Here t denotes the mass (m/z). In practice, the information over $\mu_i(t)$ is collected at a finite number of points, T_i, thus observing the data vector $\mathbf{y}_i = (y_{i1}, \ldots, y_{iT_i})^t$. The basic statistical model

Correspondence and reprint requests to Abdelali Haoudi, Eastern Virginia Medical School, Department of Microbiology and Molecular Cell Biology, Norfolk, VA 23507, USA, Email: haoudia@evms.edu

This is an open access article distributed under the Creative Commons Attribution License which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

© 2005 Hindawi Publishing Corporation

of FDA is given by

$$y_{ij} = \hat{\mu}_i(t_{ij}) = \mu_i(t_{ij}) + \epsilon_i(t_{ij}), \quad i = 1, \ldots, n, \ j = 1, \ldots, T_i, \tag{1}$$

where t_{ij} is the mass value at which the jth measurement is taken for the ith function μ_i. The independent disturbance terms $\epsilon_i(t_{ij})$ are responsible for roughness in $\mathbf{y}_i$. FDA has been developed for analyzing functional (or curve) data. In FDA, data consists of functions not of vectors. Samples are taken at time points $t_1, t_2, \ldots$, and regard $\mu_i(t_{ij})$ as multivariate observations. In this sense the original functional y_{ij} can be regarded as the limit of $\mu_i(t_{ij})$ as the sampling interval tends to zero and the dimension of multivariate observations tends to infinity. Ramsay and Silverman [12, 13] have discussed several methods for analyzing functional data, including functional regression analysis, functional principal component analysis (PCA), and functional canonical correlation analysis (CCA). These methodologies look attractive, because one often meets the cases where one wishes to apply regression analysis and PCA to such data. In the following we describe how to use the FDA tools for applying FDA and a new dissimilarity measure to classify the spectra data.

We propose to implement a hierarchical clustering algorithm for proteomics data using FDA. We use functional transformation to smooth and reduce the dimensionality of the spectra and develop a new algorithm for clustering high-dimensional proteomics data.

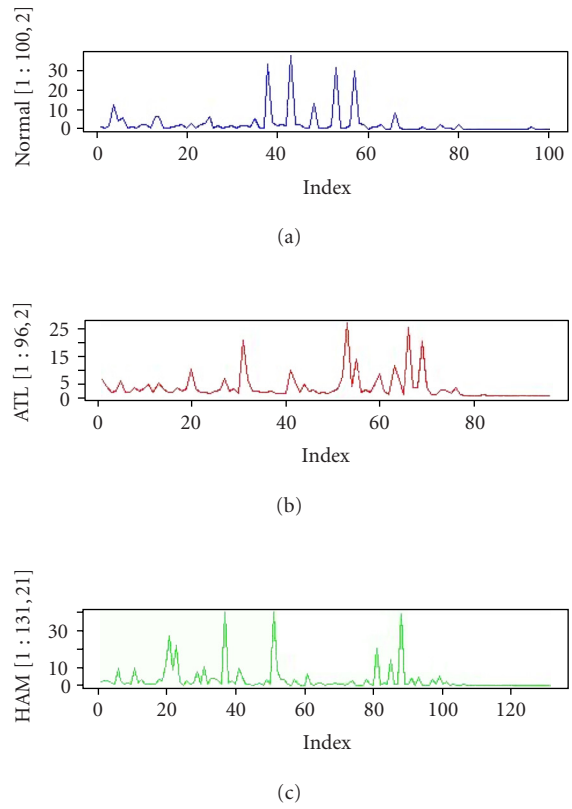

Figure 1. Three cut expressions from a normal, an HAM, and an ATL patient.

MATERIAL AND METHODS

Serum samples from HTLV-1-infected patients

Protein expression profiles generated through SELDI analysis of sera from human t-cell leukemia virus type 1- (HTLV-1)-infected individuals were used to determine the changes in the cell proteome that characterize adult T-cell leukemia (ATL), an aggressive lymphoproliferative disease from HTLV-1-associated myelopathy/tropical spastic paraparesis (HAM/TSP), a chronic progressive neurodegenerative disease. Both diseases are associated with the infection of T cells by HTLV-1. The HTLV-1 virally encoded oncoprotein Tax has been implicated in the retrovirus-mediated cellular transformation and is believed to contribute to the oncogenic process through induction of genomic instability affecting both DNA repair integrity and cell cycle progression [14, 15]. Serum samples were obtained from the Virginia Prostate Center Tissue and body fluid bank. All samples had been procured from consenting patients according to protocols approved by the Institutional Review Board and stored frozen. None of the samples had been thawed more than twice.

Triplicate serum samples ($n = 68$) from healthy or normal ($n_1 = 37$), ATL ($n_2 = 20$), and HAM ($n_3 = 11$) patients were processed. A bioprocessor, which holds 12 chips in place, was used to process 96 samples at one time. Each chip contained one "QC spot" from normal pooled serum, which was applied to each chip along with the test samples in a random fashion. The QC spots served as quality control for assay and chip variability. The samples were blinded for the technicians who processed the samples. The reproducibility of the SELDI spectra, that is, mass and intensity from array to array on a single chip (intra-assay) and between chips (interassay), was determined with the pooled normal serum QC sample (Figure 1).

SELDI mass spectrometry

Serum samples were analyzed by SELDI mass spectrometry as described earlier [16]. The spectral data generated was used in this study for the development of the novel FDA.

Hierarchical clustering using functional data analysis

We propose to implement a hierarchical clustering algorithm for proteomics data using FDA, which consists of detecting hidden group structures within a functional dataset. We apply a new dissimilarity measure to the smoothed (transformed) proteomics functions $\hat{\mu}_i$. Then

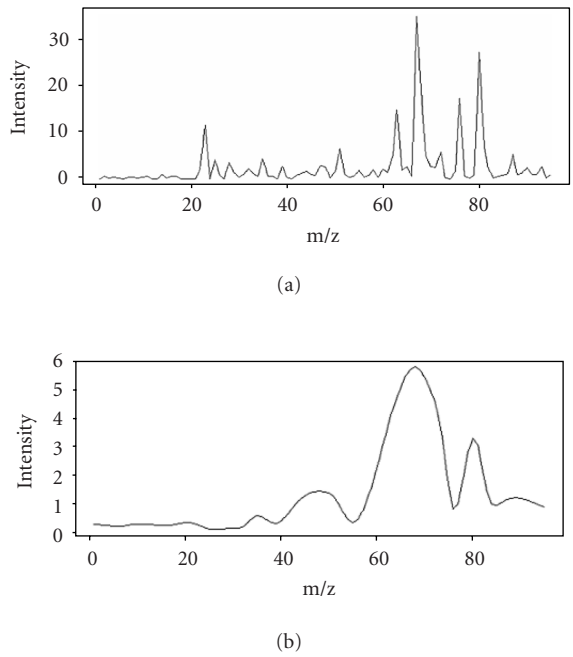

Figure 2. Original curve and a smoothed curve.

we develop a new metric that calculates the dissimilarity between different curves produced by protein expression. The development of metrics for curve and time-series models was first addressed by Piccolo [17] and Corduas [18]. Heckman and Zamar proposed a dissimilarity measure δ_{HZ} for clustering curves [19]. Their dissimilarity measure considers curve invariance under monotone transformations. Let $\Lambda_i = \{\lambda_1^{(i)}, \lambda_2^{(i)}, \ldots, \lambda_{m_i}^{(i)}\}$ be the collection of the estimated points where the curve $\mu_i(t)$ has a local maximum and let m_i be the number of maximals per observation or per sample $(i) \cdot \delta_{HZ}$ is defined as

$$\delta_{HZ}(i,l) = \frac{\sum_{j=1}^{m_i}\left(r\left(\lambda_j^{(i)}\right) - \overline{r(\lambda^{(i)})}\right)\left(r\left(\lambda_j^{(l)}\right) - \overline{r(\lambda^{(l)})}\right)}{\sum_{j=1}^{m_i}\left(r\left(\lambda_j^{(i)}\right) - \overline{r(\lambda^{(i)})}\right)^2 \sum_{j=1}^{m_l}\left(r(\lambda^{(l)}) - \overline{r(\lambda^{(l)})}\right)^2}, \tag{2}$$

where

$$r\left(\lambda_j^{(i)}\right) = k_j^{(i)} + \frac{u_j^{(i)}}{2}, \quad k_j^{(i)} = \left\{\#i,\ \lambda_i^{(i)} < \lambda_j^{(i)}\right\},$$
$$u_j^{(i)} = \left\{\#i,\ \lambda_i^{(i)} = \lambda_j^{(i)}\right\}, \qquad \overline{r(\lambda^{(i)})} = \frac{1}{m_i}\sum_{j=1}^{m_i} r\left(\lambda_j^{(i)}\right). \tag{3}$$

This measure is powerful for regression curves which are mainly monotone. On the other hand, Cerioli et al [20] propose a dissimilarity measure δ_C extending the one proposed by Ingrassia et al [21]. Cerioli's dissimilarity δ_C is defined by

$$d(i,l) = \sum_{j=1}^{m_i} \frac{|\lambda_j^{(i)} - \lambda_{*j}^{(l)}|}{m_i},$$
$$\lambda_{*j}^{(l)} = \left\{\lambda_{j'}^{(l)} : |\lambda_j^{(i)} - \lambda_{j'}^{(l)}| = \min,\ i = 1, \ldots, n\right\}, \tag{4}$$
$$\delta_C(i,l) = \left(\frac{d_{il} + d_{li}}{2}\right).$$

Both dissimilarity measures show good performance for time-series data. Dissimilarity δ_C does not involve all the indices m_i of the smoothed curve. It also uses the shortest distance between curves by involving few data points obtained by FDA smoothing.

A flexible dissimilarity measure is the one that may combine the characteristic of both measures δ_{HZ} and δ_C. This means that a potential dissimilarity measure should use the collected estimated points of the original curve obtained from FDA so that no information is lost and should work on different type of smoothed curves without using the monotonicity restriction.

In this sense, we propose a functional-based dissimilarity δ_B measure which uses the rank of the curve proposed by Heckman and Zamar and generalizes Cerioli et al dissimilarity measure as follows:

$$d_{il} = \sum_{j=1}^{m_i} \frac{\left|r\left(\lambda_j^{(i)}\right) - r\left(\lambda_{*j}^{(l)}\right)\right|}{m_i},$$
$$r\left(\lambda_{*j}^{(l)}\right) = \frac{\sum_{h=1}^{m_l}\left|r\left(\lambda_j^{(i)}\right) - r\left(\lambda_{h'}^{(l)}\right)\right|}{m_l},$$
$$r\left(\lambda_j^{(i)}\right) = k_j^{(i)} + \frac{u_j^{(i)}}{2}, \quad k_j^{(i)} = \left\{\#i,\ \lambda_i^{(i)} < \lambda_j^{(i)}\right\}, \tag{5}$$
$$u_j^{(i)} = \left\{\#i,\ \lambda_i^{(i)} = \lambda_j^{(i)}\right\}, \qquad \overline{r(\lambda^{(i)})} = \frac{1}{m_i}\sum_{j=1}^{m_i} r\left(\lambda_j^{(i)}\right).$$

Obviously, $d_{ii} = 0$ and $d_{il} = 0$, if μ_i and μ_l have the same shape $(T_i = T_l)$. We can adjust the formula above to obtain a dissimilarity measure that satisfies symmetry, by taking δ_B as our proposed dissimilarity measure:

$$\delta_B(i,l) = \left(\frac{d_{il} + d_{li}}{2}\right). \tag{6}$$

We used three powerful hierarchical methods to derive clusters or patterns using δ_B and we compare the performance of δ_B to δ_C and δ_{HZ}. The hierarchical algorithms we used are (1) *Pam* which partitions the data into different clusters "around their medoids," (2) *Clara* which works as in "Pam." Once the number of clusters is specified and representative objects have been selected from the sub-dataset, each observation of the entire dataset is assigned to the nearest medoid [22]. The sum of the dissimilarities of the observations to their closest medoid is used as a measure of the quality of the clustering. The sub-dataset for which the sum is minimal, is retained. Each sub-dataset is forced to contain the medoids obtained

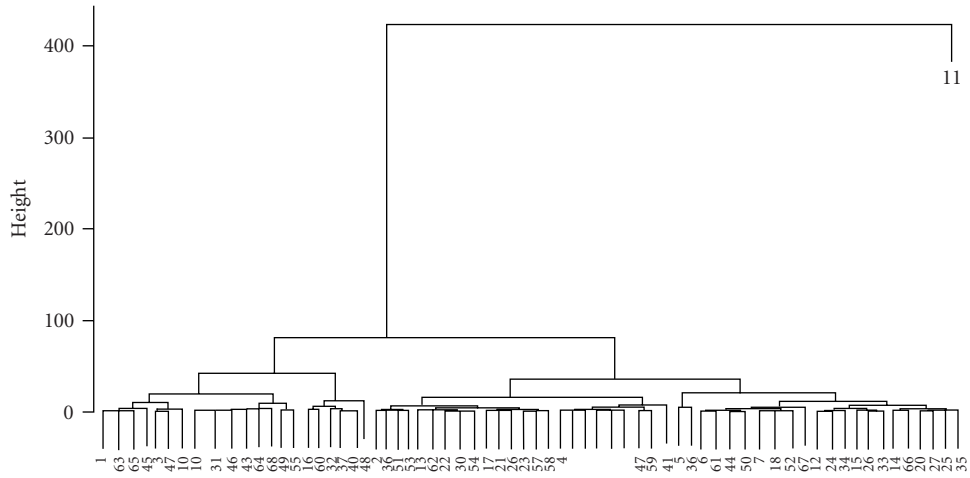

FIGURE 3. Clustering proteomics data with Diana.

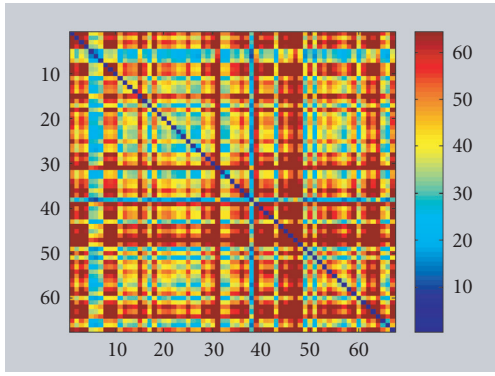

FIGURE 4. Pattern recognition using dissimilarity matrix δ_C.

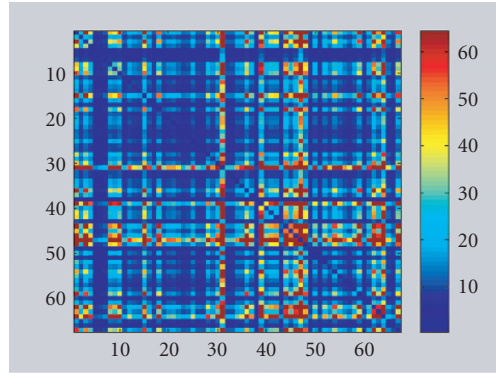

FIGURE 5. Pattern recognition using δ_{HZ}.

from the best sub-dataset until then. (3) *Diana* is probably unique in computing a divisive hierarchy, whereas most other software for hierarchical clustering is agglomerative. Moreover, Diana provides the divisive coefficient which measures the amount of clustering structure found. The Diana-algorithm constructs a hierarchy of clustering starting with one large cluster containing all n observations. Clusters are divided until each cluster contains only a single observation. At each stage, the cluster with the largest diameter is selected [22].

RESULTS

Functional data transformation reduces the dimensionality of the spectra

The spectral data were collected from proteomics analysis of a total number of serum samples ($n = 68$) including healthy or normal ($n_1 = 37$), ATL ($n_2 = 20$), and HAM ($n_3 = 11$) patients. The dataset is represented by an $n \times p$ matrix $\mathbf{X}$, where $p = 25,196$ is the number of variables (peaks) measured on each sample and $n = 68$ is the number of samples (patients). Any clustering algorithm on a datum ($68 \times 25,196$) will fail because of the singularity of the covariance matrix ($n < p$) and it will be difficult in manipulating matrices with 68 rows and 25,196 columns which has 1.7133×10^6 elements. This

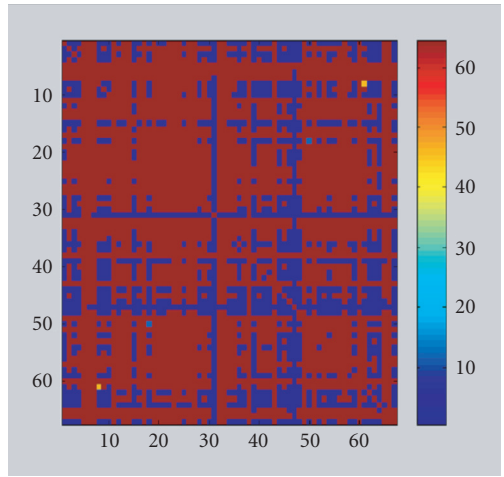

FIGURE 6. Pattern recognition using δ_B.

problem would not be raised for heuristic-based (ie, pairwise similarity-based) clustering algorithms.

To reduce the dimensionality of the spectral data, we applied FDA by fitting a P-spline curve $\widehat{\mu}_i(t)$ to each sample $\mathbf{y}_i$. P-splines satisfy a penalized residual sum of squares criterion, where the penalty involves a specified degree of derivation for $\mu_i(t)$. For example, cubic splines functions are P-splines of second order, penalizing the second derivative of $\mu_i(t)$. P-splines curves of order 3 penalize the third derivative of $\mu_i(t)$. P-splines curves of order 4 lead to an estimate of $\mu_i(t)$ with continuous first and second derivatives. We choose here to fit a P-spline curve of order 4 (Figure 2). The fitting step is performed by fixing the number of degrees of freedom that are implicit in the smoothing procedure [23].

The next step performed on the smoothed curves is to find the landmarks or indices T_i. We collected the first derivative of $\widehat{\mu}_i(t)$, say $\widehat{\mu}_i^{'}(t)$, using a smoothing P-spline function available in *R*. Those derivatives are crucial at determining the cut-off points or indices of $\mu_i(t)$. We performed this step by computing an approximate 95% pointwise confidence interval for the first derivative of $\mu_i(t)$ [24]. When the lower limit of this interval is positive, we have the confidence that $\mu_i(t)$ will be increasing. When the upper limit of this interval is negative, we have the confidence that $\mu_i(t)$ will be decreasing. Inside the interval, when the derivative changes from negative to positive, we have an optimal value which is a minimum. When the derivative changes from positive to negative, we have an optimal value which is a maximum. The maximum is set, for convenience, as the largest value of $\widehat{\mu}_i^{'}(t)$ in that interval. In this study, we restricted the choice of indices to maximal values. Let $\Lambda_i = \{\lambda_1^{(i)}, \lambda_2^{(i)}, \ldots, \lambda_{m_i}^{(i)}\}$ be the collection of the estimated points where the curve $\mu_i(t)$ has a local maximum and let m_i be the number of maximals per observation or per sample (i). Consequently, dissimilarity measure is calculated to derive the dissimilarity matrices of size $(n \times n)$ for all samples using the maximum values.

Clustering spectral data using functional data analysis

The application of functional data transformation led to the reduction of the dimensionality of the spectra to half. The size of mass indices become 12, 598. To cluster the reduced data, we calculated the three dissimilarity matrices M_{δ_C}, M_{δ_B}, and $M_{\delta_{HZ}}$. It appears that an unusual sample (patient 11) hides a possible pattern that we are trying to discover. Figure 3 shows a clustering dendrogram of the data using Diana approach. Pam and Clara gave the same results. This suggests that sample 11 would be important for further investigation.

When we removed observation 11, we detected a fewer fuzzy patterns with δ_C (Figure 4), δ_{HZ} (Figure 5), and δ_B (Figure 6). To be more specific, we investigated clusters proposed by δ_C and δ_{HZ}. A large number of clusters were proposed by both approaches (about 10 clusters). This strange result might be caused by the monotonicity assumption when using δ_{HZ} or the loss of information when using δ_C.

Table 1. Confusion matrix to show the performance of δ_B using Diana.

		Predicted			
	Classification	HAM	ATL	NOR	Total
Clinical	HAM	8	3	0	11
	ATL	5	14	1	20
	NOR	1	2	34	37
	Classification rate	0.73	0.70	0.92	0.84

Table 2. Confusion matrix to show the performance of δ_B using Clara.

		Predicted			
	Classification	HAM	ATL	NOR	Total
Clinical	HAM	10	1	0	11
	ATL	2	18	0	20
	NOR	1	1	35	37
	Classification rate	0.91	0.90	0.95	0.93

For δ_B, we provided the dendogram of the data using Diana approach (Figure 7). Three clusters were apparent. One well-separated cluster and two overlapped ones. For δ_{HZ} and δ_C, no structure was apparent which confirms the limitations of both dissimilarities as explained before.

To check the performance of our method, we calculated the confusion matrix between the predicted clusters and the clinical clusters using Diana (Table 1) and Clara (Table 2). We find that 3 patients out of 11 were misclassified for cluster 1 (HAM), 6 out of 20 were misclassified for cluster 2 (ATL), and 3 out of 37 were misclassified for cluster 3 (normal). Ham and ATL shared the majority of the misclassified observations which makes sense since both groups gather patients with a disease caused by the same retrospective virus. The error rate of misclassification for both clusters (HAM and ATL) is about 20%. For normal patient, the error rate of misclassification is about 8%. The total rate of misclassification is about 16%.

When we used Clara-based hierarchical cluster algorithm with δ_B, the classification result has dramatically been improved (Figure 8). The error rate of misclassification is reduced to 7%.The error rate of misclassification between HAM and ATL is about 9%, 5% of normal patients was misclassified. This result shows that a hierarchical δ_B dissimilarity algorithm based on minimizing the dissimilarity of observations to their closest medoid performs better than a divisive hierarchical clustering algorithm based on δ_B.

DISCUSSION

Cancer biomarkers can be used to screen asymptomatic individuals in the population, assist diagnosis in

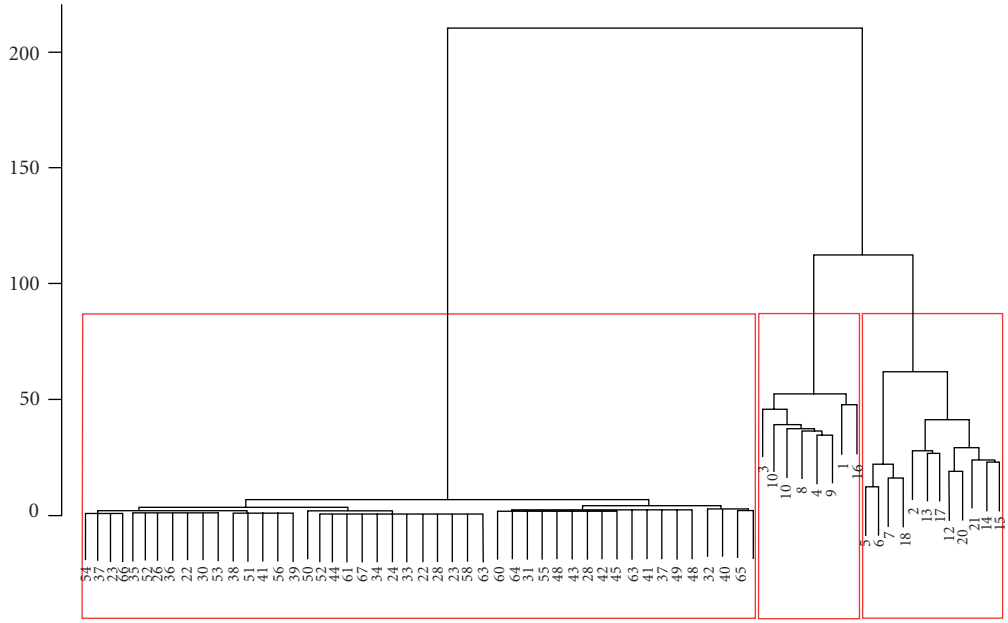

FIGURE 7. Dendogram of the δ_B dissimilarity approach with Diana.

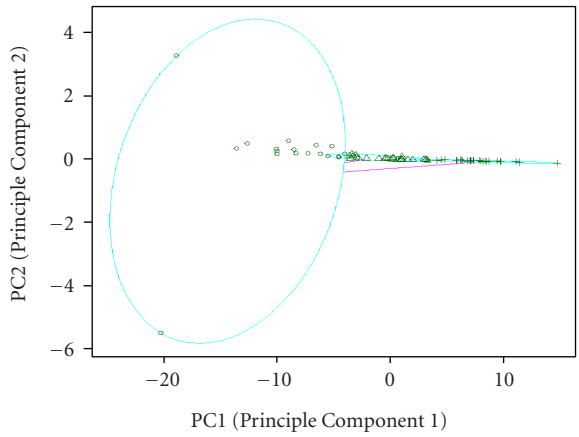

FIGURE 8. The δ_B dissimilarity approach with Clara.

suspected cases, predict prognosis and response to specific treatments, and monitor patients after primary therapy. The introduction of new technologies to the proteome analysis field, such as mass spectrometry, have sparked new interest in cancer biomarkers allowing for more effective diagnosis of cancer by using complex proteomic patterns or for better classification of cancers, based on molecular signatures, respectively. These technologies provide wealth of information and rapidly generate large quantities of data.

Processing the large amounts of data will lead to useful predictive mathematical descriptions of biological systems which will permit rapid identification of novel therapeutic targets and diseases biomarkers.

Clustering and analyzing proteomics data has been proven to be a challenging task.

Proteomics data are provided usually as curves or spectra with thousand of peaks. A clustering algorithm based on a matrix of n observations (n samples which is usually small) and p peaks (p variables which is usually a large number) will be unsuccessful. A matrix of size ($n \ll p$) will be singular and any method based on a matrix M ($n \times p$) will not be robust enough and will induce errors. A clustering algorithm based on a well-chosen dissimilarity matrix ($n \times n$) is more appropriate and more robust given the relatively moderate size of the matrix.

The use of a smoothing function for the spectra performs better for time series or for monotonic curves. We have previously successfully applied this smoothing function to large-scale proteomics data [25].

The application of Euclidean or Mahalanobis distances for instance may not perform well for this proteomics dataset, since those distances usually successfully applied to a typical data with specific expression, spherical or ellipsoidal (normally distributed data). A new dissimilarity measure has to involve other criteria such as the wealth of data points for each observation and the parallel nature expressed by the proteomics curve (or time series). On the other hand, a robust dissimilarity measure may perform badly on a curve with large data points or peaks.

Functional smoothing of proteomics expression profiles or spectra has proven to be very helpful. This has allowed us to minimize the number of peaks to retain only the ones that passed the performance of the FDA smoothing. In this study, after using FDA, we succeeded in retaining 50% of the smoothed peaks. The FDA with

the dissimilarity measure δ_B shows better performance by comparison to δ_C and δ_{HZ} known to perform well along with FDA on times-series data or on monotonic curves.

The two remaining difficulties that naturally arose are (1) to find meaningful peaks that can be used to provide better discrimination between the clusters, (2) to propose the optimal number of clusters instead of choosing them a priori. The model selection criteria might be useful to answer those questions. In fact, model selection scores use two components for selecting the number of variables and the number of clusters in a given density-based cluster analysis. The first term is the lack of fit generally proportional to the likelihood function. The second term is the penalty term (complexity term). For such proteomics dataset, we propose to use the sum of the negative δ_B dissimilarity measure between all the observations to their closest medoids as a lack of fit function. The penalty term might be simple to derive but biased using AIC and BIC, for example, or it can be more difficult to derive if one used a more robust method such as information complexity-based criteria.

ACKNOWLEDGMENT

This work was supported by the SRGP Award by the College of Business, University of Tennessee in Knoxville, by the Leukemia Lymphoma Society, and the National Institutes of Health.

REFERENCES

[1] Aebersold R, Mann M. Mass spectrometry-based proteomics. *Nature*. 2003;422(6928):198–207.

[2] Steen H, Mann M. The ABC's (and XYZ's) of peptide sequencing. *Nat Rev Mol Cell Biol*. 2004;5(9):699–711.

[3] Wright Jr GL. SELDI proteinchip MS: a platform for biomarker discovery and cancer diagnosis. *Expert Rev Mol Diagn*. 2002;2(6):549–563.

[4] Reddy G, Dalmasso EA. SELDI protein chip(R) array technology: protein-based predictive medicine and drug discovery applications. *J Biomed Biotechnol*. 2003;2003(4):237–241.

[5] Tang N, Tornatore P, Weinberger SR. Current developments in SELDI affinity technology. *Mass Spectrom Rev*. 2004;23(1):34–44.

[6] Espina V, Mehta AI, Winters ME, et al. Protein microarrays: molecular profiling technologies for clinical specimens. *Proteomics*. 2003;3(11):2091–2100.

[7] Zhang H, Yan W, Aebersold R. Chemical probes and tandem mass spectrometry: a strategy for the quantitative analysis of proteomes and subproteomes. *Curr Opin Chem Biol*. 2004;8(1):66–75.

[8] Vazquez A, Flammini A, Maritan A, Vespignani A. Global protein function prediction from protein-protein interaction networks. *Nat Biotechnol*. 2003;21(1):697–700.

[9] Bensmail H, Haoudi A. Postgenomics: proteomics and bioinformatics in cancer research. *J Biomed Biotechnol*. 2003;2003(4):217–230.

[10] Somorjai RL, Dolenko B, Baumgartner R. Class prediction and discovery using gene microarray and proteomics mass spectroscopy data: curses, caveats, cautions. *Bioinformatics*. 2003;19(12):1484–1491.

[11] Schwartz SA, Weil RJ, Johnson MD, Toms SA, Caprioli RM. Protein profiling in brain tumors using mass spectrometry: feasibility of a new technique for the analysis of protein expression. *Clin Cancer Res*. 2004;10(3):981–987.

[12] Ramsay JO, Silverman BW. *Functional Data Analysis*. New York, NY: Springer; 1997.

[13] Ramsay JO, Silverman BW. *Applied Functional Data Analysis: Methods and Case Studies*. New York, NY: Springer; 2002.

[14] Haoudi A, Semmes OJ. The HTLV-1 tax oncoprotein attenuates DNA damage induced G1 arrest and enhances apoptosis in p53 null cells. *Virology*. 2003;305(2):229–239.

[15] Haoudi A, Daniels RC, Wong E, Kupfer G, Semmes OJ. Human T-cell leukemia virus-I tax oncoprotein functionally targets a subnuclear complex involved in cellular DNA damage-response. *J Biol Chem*. 2003;278(39):37736–37744.

[16] Adam BL, Qu Y, Davis JW, et al. Serum protein fingerprinting coupled with a pattern-matching algorithm distinguishes prostate cancer from benign prostate hyperplasia and healthy men. *Cancer Res*. 2002;62(13):3609–3614.

[17] Piccolo D. A distance measure for classifying ARIMA models. *Journal of Time Series Analysis*. 1990;11:153–164.

[18] Corduas M. La metrica autoregressiva tra modelli ARIMA: una procedura in linguaggio GAUSS. *Quaderni di statistica*. 2000;2:1–37.

[19] Heckman N, Zamar R. Comparing the shapes of regression function. *Biometrika*. 2000;87(1):135–144.

[20] Cerioli A, Laurini F, Corbellini A. Functional cluster analysis of financial time series. In: *Proceedings of the Meeting of Classification and Data Analysis Group of the Italian Statistical Society (CLADAG 2003)*. Bologna, Italy: CLUEB; 2003:107–110.

[21] Ingrassia S, Cerioli A, Corbellini A. Some issues on clustering of functional data. In: Schader M, Gaul W, Vichi M, eds. *Between Data Science and Applied Data Analysis*. Berlin, Germany: Springer; 2003:49–56.

[22] Kaufman L, Rousseeuw PJ. *Finding Groups in Data. An Introduction to Cluster Analysis*. New York, NY: John Wiley & Sons; 1990.

[23] Hastie TJ, Tibshirani RJ. *Generalized Additive Models*. London UK: Chapman & Hall; 1990.

[24] Silverman BW. Some aspects of the spline smoothing approach to nonparametric regression curve fitting. *J Roy Statist Soc B*.1985;47:1–52.

[25] Bensmail H, Semmens J, Haoudi A. Bayesian fast-Fourier transform based clustering method for proteomics data. *Journal of Bioinformatics*. In press.