Classification of Proteomic Signals by Block Kriging Error Matching

Tuan D. Pham[1], Dominik Beck[1], Miriam Brandl[1], and Xiaobo Zhou[2]

[1] Bioinformatics Applications Research Center
James Cook University
Townsville, QLD 4811, Australia
[2] HCNR Center for Bioinformatics
Harvard Medical School
Boston, MA 02115, USA
tuan.pham@jcu.edu.au

Abstract. One of recent advances in biotechnology offers high-throughput mass-spectrometry data for disease detection, prevention, and biomarker discovery. In fact proteomics has recently become an attractive topic of research in biomedicine. Signal processing and pattern classification techniques are inherently essential for analyzing proteomic data. In this paper the estimation method of block kriging is utilized to derive an error matching strategy for classifying proteomic signals with a particular application to the prediction of cardiovascular events using clinical mass spectrometry data. The proposed block kriging based classification technique has been found to be superior to other recently developed methods.

Keywords: Proteomics, mass spectral data, block kriging, signal processing, classification, distortion measures.

1 Introduction

The study of proteomic patterns have recently been utilized for early detection of disease progressions [1,2,3]. Mass spectrometry data has been playing a major role in the discovery of disease pathways and biomarkers for new drug treatment and development [4,5].

Methods for classification of normal and cancerous states using mass spectrometry (MS) data have been recently developed. Petricoin *et al.* [2] applied cluster analysis and genetic algorithms to detect early stage ovarian cancer using proteomic spectra. Ball *et al.* [6] applied integrated approach based on neural networks to study SELDI-MS data for classification of human tumors and identification of biomarkers. Lilien *et al.* [7] applied principal component analysis and a linear discriminant function to classify ovarian and prostate cancers. Sorace and Zhan [8] used mass spectrometry serum profiles to detect early ovarian cancer. Wu *et al.* [9] compared the performance of several methods for the classification of mass spectrometry data. Tibshirani *et al.* [10] proposed a probabilistic approach for sample classification from protein mass spectrometry data. Morris

A. Elmoataz et al. (Eds.): ICISP 2008, LNCS 5099, pp. 281–288, 2008.

et al. [11] applied wavelet transforms and peak detection for feature extraction of MS data. Yu *et al.* [12] developed a method for dimensionality reduction for high-throughput MS data. Levner [13] used feature selection methods and then applied the nearest centroid technique to classify MS-based ovarian and prostate cancer datasets.

Given the promising integration of several classification methods and mass spectrometry data in high-throughput proteomics [14], this new biotechnology still encounters several challenges in order to become a mature platform for clinical diagnostics and protein-based biomarker profiling. One of a major concerns is the finding of an effective computational approach for the analysis of this type of high-throughput data. In this paper we discuss for the first time the implementation of blocking kriging technique to determine the long-range spatial error variances of mass spectrometry data that can be used as a basis for signal matching and classification.

2 Error Matching by Block Kriging

Kriging techniques [15,16] estimate the unknown value at a particular location as the linear combination of the known values at nearby locations:

$$\hat{s}(n) = \sum_{k=1}^{p} w_k s(n_k) \tag{1}$$

where w_k, $k = 1, \ldots, p$ are the kriging weights, and $s(n_k)$, $k = 1, \ldots, p$, are the known data values at locations n_k.

The central idea to this estimation is to determine a set of optimal kriging weights. Using the method of block kriging, these optimal weights can be obtained as

$$\mathbf{C}\mathbf{w} = \mathbf{b} \tag{2}$$

where $\mathbf{C}$ is the square and symmetrical matrix that represents the spatial covariances between the known signals, $\mathbf{w}$ is the vector of kriging weights and a Lagrange multiplier μ, and $\mathbf{b}$ is the vector that represents the average spatial covariances between a particular sample location and all the points within a domain A:

$$\mathbf{C} = \begin{bmatrix} C_{11} & \cdots & C_{1n} & 1 \\ . & \cdots & . & . \\ . & \cdots & . & . \\ . & \cdots & . & . \\ C_{n1} & \cdots & C_{nn} & 1 \\ 1 & \cdots & 1 & 0 \end{bmatrix}$$

$$\mathbf{w} = \begin{bmatrix} w_1 \cdots w_n \; \mu \end{bmatrix}^T$$

$$\mathbf{b} = \begin{bmatrix} C_{1A} \cdots C_{nA} \; 1 \end{bmatrix}^T$$

The sample spatial covariance used for the kriging estimator can be calculated as

$$C_{ij} = \frac{1}{N(h)} \sum_{(i,j)|h_{ij}=h} s(j) - (\frac{1}{n}\sum_{k=1}^{n} s(k))^2 \tag{3}$$

in which the sample spatial covariance is a function of the lag distance h, $N(h)$ is the number of pairs that $s(i)$ and $s(j)$ are separated by h, and n is the total number of data points.

The average spatial covariances between a sample location and all the points within A is defined as

$$C_{iA} = \frac{1}{N} \sum_{j|j \in A}^{N} C_{ij} \tag{4}$$

Thus the vector of the spatial predictor coefficients can be obtained by solving

$$\mathbf{w} = \mathbf{C}^{-1}\,\mathbf{b} \tag{5}$$

The block kriging error variance is given by

$$\sigma^2_{BK} = C_{AA} - \mathbf{w}^T\mathbf{b} \tag{6}$$

where

$$C_{AA} = \frac{1}{NM} \sum_{i|i \in A}^{M} \sum_{j|J \in A}^{N} C_{ij} \tag{7}$$

In terms of the semi-variogram, the block kriging error variance is given by

$$\sigma^2_{BK} = \mathbf{w}^T\mathbf{b} \tag{8}$$

where all the covariance terms involving in $\mathbf{C}$ and $\mathbf{b}$ expressed in (2) are replaced with the semi-variogram values.

The semi-variogram is a function which expresses the spatial relationship of a regionalized variable [17]. In probabilistic notation, the variogram, $2\gamma(h)$, is defined as the expected value:

$$2\gamma(h) = E\{[s(i) - s(j)]^2\},\ h_{ij} = h \tag{9}$$

where h is a lag distance that separates $s(i)$ and $s(j)$.

The semi-variogram is half of the variogram, that is, $\gamma(h)$. The experimental semi-variogram for lag distance h is defined as the average squared difference of values separated by h:

$$\gamma(h) = \frac{1}{2N(h)} \sum_{(i,j)|h_{ij}=h} [s(i) - s(j)]^2 \tag{10}$$

where $N(h)$ is the number of pairs for lag h.

We now consider $\mathbf{x}$, $\mathbf{y}$, and $\mathbf{z}$ as the vectors defined on a vector space V. A metric or distance d on V is defined as a real-valued function on the Cartesian product $V \times V$ if it has the properties of positive definiteness, symmetry, and triangle inequality. If a measure of dissimilarity satisfies only the property of positive definiteness, it is referred to as a distortion measure which is considered very common for the vectorized representations of signal spectra [18]. In general, to calculate a distortion measure between two vectors $\mathbf{x}$ and $\mathbf{y}$, denoted as $D(\mathbf{x}, \mathbf{y})$, is to calculate a cost of reproducing any input vector $\mathbf{x}$ as a reproduction of vector $\mathbf{y}$. Given such a distortion measure, the mismatch between two signals can be quantified by an average distortion between the input and the final reproduction. Intuitively, a match of the two patterns is good if the average distortion is small.

A very useful distortion measure that is derived from the above mathematical basis is the likelihood ratio distortion between the two templates presented in the form of two vectors of predictor coefficients $\mathbf{w}$, and $\mathbf{w}'$ which are used to model signal s. The likelihood-ratio distortion measure, denoted by D_{LR}, is defined as [18]

$$D_{LR}(\mathbf{w}, \mathbf{w}') = \frac{\mathbf{w}'^T \, \mathbf{R}_s \, \mathbf{w}'}{\mathbf{w}^T \, \mathbf{R}_s \, \mathbf{w}} - 1 \tag{11}$$

where $\mathbf{R}_s$ is the autocorrelation matrix of sequence s associated with its prediction coefficient vector $\mathbf{w}$, and $\mathbf{w}'$ is the prediction coefficient vector of signal s'. For a perfect match between the two templates, the errors are identical and (11) yields a zero distortion. For a mismatch, the residual resulting from the prediction analysis is large and the distortion defined in (11) therefore becomes large.

Based on the same principle derived for the likelihood ratio distortion, the block kriging distortion measure, denoted as D_{BK}, can be defined as

$$D_{BK}(\mathbf{w}, \mathbf{w}') = \frac{\mathbf{w}'^T \mathbf{b}}{\mathbf{w}^T \mathbf{b}} - 1 \tag{12}$$

where $\mathbf{w}$ is defined in (2) which is the kriging vector of signal s, $\mathbf{b}$ is the vector, in terms of the semi-variogram values, defined in (2) associated with s, and $\mathbf{w}'$ is the kriging vector of signal s'.

If the input (unknown) MS signal s_m is analyzed by the prediction analysis which results in a set of prediction coefficients, then the spectral distortion between an unknown sample s_m and a particular known class i can be determined using the minimum rule as follows.

$$D_{min}(\mathbf{x}_m, \mathbf{c}^i) = \min_j D(\mathbf{x}_m, \mathbf{c}^i_j) \tag{13}$$

where D is a spectral distortion measure, $\mathbf{x}_m$ is the prediction vector of s_m, $\mathbf{c}^i_j$ is the prediction vector of the j sample that belongs to class i.

Using a simple decision logic, the unknown signal s_m is assigned to class i^* if the minimum distortion measure of its prediction vector $\mathbf{x}_m$ and the corresponding prediction vector $\mathbf{c}^i$ is minimum, that is

$$s_m \rightarrow i^*, \;\; i^* = \arg\min_i D_{min}(\mathbf{x}_m, \mathbf{c}^i) \tag{14}$$

3 Experimental Results

We used high-throughput, low-resolution SELDI MS (www.ciphergen.com) to acquire the protein profiles from patients and controls. The protein profiles were acquired from 2 kDa to 200kDa. The design of the experiment originally described in [19] involves the datasets for the control and MACE group.

For the control group, the dataset consists of sixty patients who presented in emergency room with chest pain and the patients' troponin T test was consistently negative. These patients lived in the next 5 years without any major cardiac events or death. The total 166 plasma samples, 24 reference samples and 6 blanks were fractionated into 6 fractions using two 96-well plates containing anion exchange resin (Ciphergen, CA).

For the MACE group, the dataset was designed to comprise 60 patients who presented in emergency room with chest pain but the patients' troponin T test was negative. However, the patients in this group had either a heart attack, died or needed revascularization in the subsequent 6 months. The blood samples used in this study were same as those used in [20]. Most new MPO data measured with FDA approved CardioMPO kit for these two groups are available – MPO levels for 56 (out of 60) patients in control group and 55 (out of 60) patients in MACE group are available. Statistical analysis shows that MPO alone can distinguish MACE from control with accuracy of better than 60%.

For the SELDI mass spectra, the coverage of proteins in SELDI protein profiles was increased by that the blood samples were fractionated with HyperD Q (strong ion exchange) into 6 fractions. The protein profiles of fractions were acquired with two SELDI Chips: IMAC and CM10. There are a few different SELDI chips with different protein binding properties. Generally speaking, the more types of the SELDI chips are used, the more proteins are likely to be detected. However, due to the high concentration dynamic range of the proteins in

Table 1. Average MACE prediction rates by different methods

Method	Average accuracy (%)
MPO value	55.25
T-test	62.23
Standard genetic algorithm	69.05
Sequential forward floating search	71.92
Improved genetic algorithm	75.16
Block kriging distortion measure	93.15

human blood, the total number of proteins to be detected by the protocol we are using is very limited. We estimate that the number of the proteins we are able to detect is about one-thousand, while the total protein number in human blood is estimated to be tens of thousands. For example, MPO can be accurately measured with immunoassay (CardioMPO) but could not be detected with SELDI MS. The MS data for each sample in each fraction was acquired in duplicate, so 120 samples (60 controls and 60 MACEs) in each fraction in one type of SELDI chip have 240 spectra. There are two types of SELDI chips: IMAC and CM10.

To emphasize our study on rigorous prediction using SELDI mass spectra, we used only two fractions to carry out the experiment. Because of the short length of the samples, we concatenated the corresponding samples of the two fractions for the extraction of the prediction coefficients. We have recently applied the statistical and geostatistical prediction models, and the prediction based classification rule for extracting the MS features and classifying control and MACE samples, respectively [21]. The pattern matching using block kriging error matching defined in (12) was carried out for the same dataset in this study.

In previous work [21], we performed the leave-one-out validation and obtained the average classification rate of 83.34% using the statistical distortion measure; whereas for the ordinary kriging distortion measure, the average classification result was 97.10%. However, these classification results were based on the use of whole MS sequences and not based on the MS peak values. In another previous work [19], we used the MPO value, five selected biomarkers by T-test, five selected biomarkers by the sequential forward floating search (SFFS), five selected biomarkers by standard genetic algorithm (GA), and five selected biomarkers by an improved genetic algorithm (IGA) to carry out the prediction. Using the same dataset and test design, the block kriging gave the average classification result as 93.15%. The average validation results of different methods are presented in Table 1 which shows the superior performance of the block kriging approach over other relevant models.

4 Conclusion

It has been predicted that the advancement of proteomics pattern diagnostics might represent a revolution in the field of molecular medicine, because this technology has the potential of developing a new model for early disease detection [3,22,23]. Given that the research into clinical proteomic pattern diagnostics is still in its infancy because the results have not been validated in large trials, and effective computational methods have not been well-explored; recent research outcomes have illustrated the role of MS-based proteomics as an indispensable tool for molecular and cellular biology and for the emerging field of systems biology [5].

Early disease detection using MS data is a challenging task up to date. This task requires the combination of the contrast fields of knowledge of modern biology and computational methodology. We have presented in this paper the novel applications of the theories of linear prediction in time and space domains for extracting the effective features of mass spectrometry data that can be useful for the classification of MS spectra. The initial results using the cardiovascular SELDI-MS datasets have shown the potential application of the proposed techniques for predicting patient's major adverse cardiac risk that can be helpful for the development of the diagnostic kit for treatment of patients during their initial emergency admission [24]. The use of the prediction coefficients can be extended to other spectral coefficients derived from the prediction principle, and the distortion measures can be used as similarity measures for other classification techniques.

Research into MS-based disease detection has recently attracted the attention of researchers from various disciplines. In particular, it offers tremendous potentials for the discovery of novel biomerkers and the development of personalized medicine [25] – a new concept that major diseases have a genetic component; therefore the understanding of cellular processes at the molecular level will enable scientists and physicians to predict the relative risk and potential therapy for such conditions on a person-to-person basis.

Acknowledgments

This research was supported under the Australian Research Council's Discovery Projects funding scheme (project number DP0877414) to Tuan Pham.

References

1. Sauter, E., et al.: Proteomic analysis of nipple aspirate fluid to detect biologic markers of breast cancer. Br. J. Cancer 86, 1440–1443 (2002)
2. Petricoin, E.F., et al.: Use of proteomic patterns in serum to identify ovarian cancer. Lancet 359, 572–577 (2002)
3. Conrads, T.P., Zhou, M., Petricoin III, E.F., Liotta, L., Veenstra, T.D.: Cancer diagnosis using proteomic patterns. Expert Rev. Mol. Diagn. 3, 411–420 (2003)
4. Griffin, T., Goodlett, T., Aebersold, R.: Advances in proteomic analysis by mass spectrometry. Curr. Opin. Biotechnol. 12, 607–612 (2002)
5. Aebersold, R., Mann, M.: Mass spectrometry-based proteomics. Nature 422, 198–207 (2003)
6. Ball, G., Mian, S., Holding, F., Allibone, R.O., Lowe, J., Ali, S., Li, G., McCardle, S., Ellis, I.O., Creaser, C., Rees, R.C.: An integrated approach utilizing artificial neural networks and SELDI mass spectrometry for the classification of human tumours and rapid identification of potential biomarkers. Bioinformatics 18, 395–404 (2002)
7. Lilien, R.H., Farid, H., Donald, B.R.: Probabilistic disease classification of expression-dependent proteomic data from mass spectrometry of human serum. J. Computational Biology 10, 925–946 (2003)
8. Sorace, J.M., Zhan, M.: A data review and re-assessment of ovarian cencer serum proteomic profiling. BMC Bioinformatics 4, 24 (2003)
9. Wu, B., Abbott, T., Fishman, D., McMurray, W., Mor, G., Stone, K., Ward, D., Williams, K., Zhao, H.: Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data. Bioinformatics 19, 1636–1643 (2003)

10. Tibshirani, R., Hastie, T., Narasimhan, B., Soltys, S., Shi, G., Koong, A., Le Sample, Q.-T.: classification from protein mass spectrometry, by peak probability contrasts. Bioinformatics 20, 3034–3044 (2004)
11. Morris, J.S., Coombes, K.R., Koomen, J., Baggerly, K.A., Kobayashi, R.: Feature extraction and quantification for mass spectrometry in biomedical applications using the mean spectrum. Bioinformatics 21, 1764–1775 (2005)
12. Yu, J.S., Ongarello, S., Fiedler, R., Chen, X.W., Toffolo, G., Cobelli, C., Trajanoski, C.Z.: Ovarian cancer identification based on dimensionality reduction for high-throughput mass spectrometry data. Bioinformatics 21, 2200–2209 (2005)
13. Levner, I.: Feature selection and nearest centroid classification for protein mass spectrometry. BMC Bioinformatics 6, 68 (2005)
14. Shin, H., Markey, M.K.: A machine learning perspective on the development of clinical decision support systems utilizing mass spectra of blood samples. J. Biomedical Informatics 39, 227–248 (2006)
15. Matheron, G.: The theory of regionalized variables and its applications. Paris School of Mines Publication, Paris (1971)
16. Isaaks, E.H., Srivastava, R.M.: An Introduction to Applied Geostatistics. Oxford University Press, New York (1989)
17. Deutsch, C.V.: Geostatistical Reservoir Modeling. Oxford University Press, New York (2002)
18. Rabiner, L., Juang, B.-H.: Fundamentals of Speech Recognition. Prentice Hall, New Jersey (1993)
19. Zhou, X., Wang, H., Wang, J., Hoehn, G., Azok, J., Brennan, M.L., Hazen, S.L., Li, K., Wong, S.T.C.: Biomarker discovery for risk stratification of cardiovascular events using an improved genetic algorithm. In: Proc. IEEE/NLM Int. Symposium on Life Science and Multimodality, pp. 42–44 (2006)
20. Brennan, M.-L., Penn, Van Lente, M.S., Nambi, V., Shishehbor, M.H., Aviles, R.J., Goormastic, M., Pepoy, M.L., McErlean, E.S., Topol, E.J., Nissen, S.E., Hazen, S.L.: Prognostic value of myeloperoxidase in patients with chest pain. The New England Journal of Medicine 13, 1595–1604 (2003)
21. Pham, T.D., Wang, H., Zhou, X., Beck, D., Brandl, M., Hoehn, G., Azok, J., Brennan, M.-L., Hazen, S.L., Li, K., Wong, S.T.C.: Computational prediction models for early detection of risk of cardiovascular events using mass spectrometry data. IEEE Transactions on Information Technology in Biomedicine (in print)
22. Petricoin, E.F., Liotta, L.A.: Mass spectrometry-based diagnostics: The upcoming revolution in disease detection. Clinical Chemistry 49, 533–534 (2003)
23. Wulfkuhle, J.D., Liotta, L.A., Petricoin, E.F.: Proteomic applications for the early detection of cancer. Nature 3, 267–275 (2003)
24. Goodacre, S., Locker, T., Arnold, J., Angelini, K., Morris, F.: Which diagnostic tests are most useful in a chest pain unit protocol? BMC Emergency Medicine 5, 6 (2005)
25. Ginsburg, G.S., McCarthy, J.J.: Personalized medicine: revolutionizing drug discovery and patient care. Trends Biotechnol. 19, 491–496 (2001)