ProteinChip® Clinical Proteomics: Computational Challenges and Solutions

Eric T. Fung and Cynthia Enderwick
CIPHERGEN Biosystems, Fremont, CA, USA

Computational Proteomics Supplement 32:S34-S41 (March 2002)

ABSTRACT

ProteinChip® technology, a suite of analytical tools that includes retentate chromatography, on-chip protein characterization, and multivariate analysis, allows researchers to examine patterns of protein expression and modification. Based on the surface enhanced laser desorption/ionization time-of-flight mass spectrometry (SELDI-TOF-MS) approach (17), ProteinChip technology has been pioneered by researchers at Ciphergen Biosystems (Fremont, CA, USA), as well as by users of Ciphergen's commercial embodiment of this technology, the ProteinChip Biomarker System. This report will begin with a background of the technology and describe its applications in clinical proteomics and will then conclude with a discussion of tools and strategies to mine the large amounts of data generated during the course of a typical clinical proteomics study.

TECHNOLOGY BACKGROUND

Because biological systems possess a large dynamic range of protein expression, proteomics techniques generally consist of one or more up-front sample fractionation procedures to simplify the proteome into sub-proteomes, followed by a detection methodology with adequate sensitivity to scan the sub-proteomes. Using ProteinChip technology, sample fractionation is accomplished by retentate chromatography while detection is accomplished by SELDI-TOF-MS. Retentate chromatography is performed on ProteinChip Arrays with varying chromatographic properties, e.g., anion exchange, cation exchange, metal affinity, and reverse phase. By utilizing arrays with differing surface chemistries in parallel and in series, a complex mixture of proteins, as from cells or body fluids, can be resolved into subsets of proteins with common properties. After the arrays are washed to remove weakly bound proteins, a solution containing an energy-absorbing molecule (often referred to as matrix) is added and allowed to crystallize, embedding the retained proteins. After crystallization, features within these arrays are read in a ProteinChip Reader. Once a peak of interest has been detected, the analyte can be enriched or purified for further analysis, which is accomplished through a combination of column and on-chip purification strategies. Once a protein is sufficient-

ly purified, it is digested with proteolytic enzymes, and subsequent analysis of the peptide patterns by single mass spectrometric means can yield important identification information, such as protein identification by peptide mass fingerprinting. Alternatively, protein mixtures upon these arrays may be studied by combining on-chip chemical (26) or enzymatic (17) digestion with tandem mass spectrometry to identify proteins using database-correlation approaches. In addition, the platform can be used to identify proteins purified from one-dimensional (1-D) or 2-D gels (for an example, see Reference 25). This article discusses strategies for protein profiling using ProteinChip technology. Various applications of SELDI-TOF-MS have been described in several reviews (9,10,17) while more technical information can be found in Reference 27.

CLINICAL PROTEOMICS

Clinical proteomics aims to scan the realm of expressed proteins to identify biomarkers that can answer specific clinical questions. The most obvious are markers that can be used for diagnosis or prognosis. Another important issue that clinical proteomics promises to help resolve includes the ability to predict a patient's response to a specific drug. Clinical proteomics also has utility outside of the doctor's office. For example, diagnostic markers can themselves be candidates for drug targets. Pharmaceutical companies pursue clinical proteomics to identify markers that predict toxicity of candidate drugs.

The most straightforward approach to clinical proteomics begins with protein profiling or protein differential display studies, in which protein expression in normal samples is compared with that in perturbed samples. Examples of perturbation can be a disease state or treatment with a candidate drug. Physiological fluids, tissue homogenates, or cell lysates from disease and control samples are processed on the same types of array surfaces, and the arrays read under the same data collection conditions. The premise of this approach is to establish composite fingerprint profiles of both disease and non-disease states from a series of training samples, and then to use these profiles to make a diagnosis on actual unknown patient samples. Biomarker identification is not

strictly required for diagnostic purposes but is required for insight into the underlying biology as well as to assess the use of the biomarker as a drug target. Methods by which one determines whether a peak satisfies statistical requirements for a biomarker will be discussed in greater detail below. Biomarker discovery efforts have been performed in a variety of diseases including infectious and inflammatory diseases, neuropsychiatry, cardiovascular disease, and cancer, and details and results of some of these profiling studies can be found elsewhere (10,13,15,23,24).

The typical clinical proteomics study we perform in Ciphergen's Biomarker Centers begins with a discovery phase, in which assay conditions are tested on a relatively small number of samples. The number of samples and types of samples are perhaps the most important parameters that determine the success of a project. Usually we profile at least 30 samples in each classification group (e.g., disease versus healthy or treated versus untreated). This number of samples is usually enough to give us >90% statistical confidence in single markers with P values $< .01$ and is also enough to allow us to use some forms of multivariate analysis. The samples themselves are another critical parameter. Because this sample set size is relatively small, inherent biological variability always threatens the ability to conclude that differences seen are consequences specific to the disturbance under study. Therefore, it is imperative that the study include well-chosen samples (e.g., patients of the same age group or all of a single sex) and, equally important, appropriately chosen controls. Naturally, *in vitro* studies show less variability than do animal studies, which in turn show less variability than human studies. No matter the source of samples, all should have been handled identically, and care should have been taken to minimize the number of freeze-thaw cycles.

Once the samples are procured, they can be processed directly on the arrays with minimal preparation. However, when there is adequate material (for serum, 20 μ L of volume), we typically perform some form of fractionation prior to any ProteinChip Array binding procedure. From our experience, this fractionation step significantly increases the number of peaks visualized and therefore increases the likelihood that biomarkers will be found (Figure 1). For serum, we use anion exchange fractionation, while for cells or tissues we often perform subcellular fractionation. Specific applications might make other types of fractionation worthwhile. For example, if one is specifically interested in assaying calcium-binding proteins, immobilized metal affinity chromatography (IMAC) fractionation should be performed. Following fractionation, each fraction is profiled under a series of ProteinChip Array assay conditions, which can include different permutations of array surface chemistries, choice of energy absorbing molecules, and laser energies. Consequently, each sample in the study generates multiple spectra and therefore data analysis and reduction must be manageable while taking advantage of the plethora of information made available.

INTRODUCTION TO COMPUTATIONAL METHODS

Analysis of ProteinChip data consists of several pre-processing and post-processing steps. We define pre-processing as manipulations required to organize the data, such as mass calibration and intensity normalization. Post-processing consists of

using analytical tools so that one can draw conclusions regarding the data. These steps are similar in many ways to those used to process DNA microarray images. For example, similar to extracting the gene expression levels from each spot in a DNA microarray (i.e., "feature extraction") (7,14,21), we must process the mass spectra in a way that is conducive to such downstream multidimensional methods such as clustering and classification. With this goal in mind, the following paragraphs outline the analysis procedure we employ during the course of a protein expression profiling experiment.

Analogies to DNA Microarrays

The spectrum acquired from a unique sample run on a single ProteinChip "spot" is analogous to the image acquired from the hybridization of a unique sample on an entire DNA microarray. Just as multiple DNA microarrays may be necessary to profile the entire genome, multiple combinations of sample fraction and ProteinChip Array surface type may be necessary to sufficiently profile the proteome. For ease of discussion, we will consider the case of profiling based on just one combination of fraction and surface type in the following section. There are two basic types of DNA arrays—the Affymetrix GeneChip[®], and the class of spotted microarrays including cDNA, *in situ* synthesized, or whole oligonucleotide arrays. Because we do not have the equivalent of a two-dye competitive hybridization as described in Reference 12, we are more similar to the GeneChip platform and deal with peak intensities, not in ratios or log ratios thereof. Perhaps the most obvious difference is the form of the raw data. While the DNA microarrays have position-dependent sequence-specific binding, the binding to ProteinChip spots used for general profiling is specific only to entire classes of proteins that share physical or chemical properties which create an affinity for a given ProteinChip Array surface. (Single antigen-specific antibody-coated array applications are available as well on the ProteinChip technology platform, but are not the focus of this paper; for examples of those applications, see References 1, 6, and 28). Consequently, mass spectra can contain hundreds of protein expression levels encoded in their peaks.

Processing the Raw Spectra (Pre-Processing)

There are a few steps in processing the mass spectra that can be done in isolation on a per-spectrum basis. The first is mass calibration, or the conversion of the raw time of flight (TOF) data to molecular weight (MW). Typically this process involves acquiring a spectrum from a standard with at least 5 proteins or peptides of various MWs, spanning the MW range of interest if possible. A quadratic equation relating the TOF to MW is then fit to the TOF values of the standard peaks in this spectrum. The equation generated by this process can then be used on mass spectra that are collected under the same instrument conditions, including laser intensity, approximate date, and focusing mass or time lag.

The second processing step that can be performed on a spectrum in isolation is baseline subtraction, which eliminates any baseline signal caused mostly by chemical noise from matrix molecules, similar to background subtraction in microarrays. Like typical microarray background subtraction that subtracts some estimate of the local or global background isolated from

PRODUCT & TECHNOLOGY REPORT

the array of features, our goal is to eliminate a baseline signal estimated without the contamination of true protein or peptide peaks. We employ a varying-width segmented convex hull algorithm to subtract the baseline; in essence, it is the equivalent of local background subtraction. The result is a spectrum with a baseline signal hovering slightly above zero with protein peaks maintaining their true intensity as in Figure 2, a and b.

Finally, peak detection is of course a key component in identifying and quantifying protein peaks in the mass spectra. The peak detection algorithm we employ identifies areas of the mass spectrum as peaks by comparing the signal to a neighboring valley depth calculation. As the signal rises out of the valley and descends back into it, that area of the spectrum is identified as a peak. A centroid algorithm is then used to find the apex of this area. Peak intensity is used to represent the relative quantity of protein expressed in the sample. The peak area calculation contains more uncertainty since it is not always possible to deconvolute the single peak of interest from the area of spectrum that may be contaminated by other peaks in the vicinity of the MW of interest.

Signal-to-noise can then be calculated for each peak and used as a filter in further processing. Noise is calculated locally based on the standard deviation from a linear regression of the signal around the point of interest. By using only a fraction of the lower intensity data points, this method filters out contaminating signal due to other peaks to prevent any skew in the noise calculation.

Multi-Spectra Processing

While we do not have the equivalent (or luxury) of dye normalization as in two-dye competitively hybridized microarrays, we do have a problem similar to inter-array normalization, or, in our case, inter-spectrum normalization (30). Normalization is essential to eliminate any systematic effects between samples due to varying amounts of protein or degradation over time in the sample or variation in the instrument detector sensitivity. We currently advocate a normalization method using the total ion current, or “total area under the curve.” In employing this method, we assume that on average, the number of proteins that are being overexpressed is approximately equal to the number of proteins being underexpressed and that the number of proteins whose expression levels change is few relative to the total number of proteins bound to the ProteinChip Array surface. The total ion current is calculated and divided by the number of points over which it was calculated, thus resulting in an average ion current. After an overall average of the ion current is calculated across all the spectra in the study, each spectrum is simply multiplied by a constant factor, which is equal to the overall average ion current divided by the average ion current for that spectrum. Internal studies on replicates of the same sample run on a number of spots of the same ProteinChip Array surface type have shown improvements in the coefficient of variation (CV) on the order of 15%–20%.

Figure 1. Fractionation increases the number of peaks visualized. Liver lysate was applied directly to a weak cation exchange ProteinChip Array (top spectrum) or fractionated on an anion exchange column and fractions taken in a stepwise pH gradient and each fraction applied to the weak cation exchange array (bottom four spectra). Several peaks are obviously seen in the fractionated material that are not evident in the unfractionated material.

It is a natural inclination to add control peptides or proteins to aid in normalization. While easily done with microarrays where spiking in a control gene does not affect the gene expression log ratios of the experimental genes (although one would not necessarily use spiked-in controls for normalization), spiking of calibration standards in ProteinChip experiments does have an effect on the resulting spectra due to a number of factors, including ion suppression, competition for surface binding, and the introduction of calibrant signal in the resulting spectra that must be deconvoluted.

Quantifying gene expression information from microarrays is primarily an image-processing problem and is straightforward in the sense that each gene of interest is located at a specific location in a regular grid of small spots or rectangles. Extracting this information from the ProteinChip data is more involved since we do not know a priori which proteins will be captured on a particular surface type from a particular sample type. However, for a group of spectra comprising a profiling study, we can leverage the information across all the spectra to find the union of all possible proteins peaks in the study, as shown in Figure 2c.

This process involves a number of different steps. The first is to perform an initial pass of peak detection for each spectrum. For studies containing spectra from different samples, this step inevitably will find a different set of protein peaks per spectrum.

The second step groups peaks of similar molecular weight across all the spectra together into peak clusters while allowing for slight variations in mass. Each cluster therefore represents a particular protein. The peaks used to generate these clusters must meet a minimum signal-to-noise ratio. At this point, each cluster will be incomplete. For example, if a population containing normal and cancer samples are contained in the study, a cancer marker may be expressed in only a sub-population of the spectra. Thus, the cluster created by this marker will have representation from the subgroup, but not from the other spectra without the marker.

The remaining missing peak intensity values in each cluster must now be filled in. In order to complete these clusters, a second, less stringent signal-to-noise ratio threshold is used to allow additional peaks to fill in the clusters. By lowering the signal-to-noise threshold, peaks that may have missed the cutoff used when forming the clusters may be included in this step. For spectra that do not have membership in a particular cluster, if a peak is found within the mass window of that cluster, it will be added to that cluster. After this second pass peak detection, if a cluster is still not complete, the peak intensity will be estimated by using the intensity at the average molecular weight of the existing members in the cluster.

At this point, we have reached our goal of quantifying the expression level for each protein present in our population of sam-

Figure 2. ProteinChip data processing steps. (A) An unprocessed spectrum showing where the baseline lies. (B) After baseline subtraction and calibration, spectra can be grouped for further analysis. (C) Once spectra of similar experimental conditions (same fraction, same surface type) are grouped, they can be normalized and peaks can be detected using an iterative scheme. (D) We use log plots and box-and-whisker plots to view the protein expression data in order to help identify potential biomarkers.

PRODUCT & TECHNOLOGY REPORT

Figure 3. Classification tree example. (A) Each node (blue square) is a decision point. Each sample is sifted down the tree based on how it answers the question in each node. For example, the first node asks the question, “Does peak at molecular weight 15 153 Da have a peak intensity < 5.169?” If the answer is yes, the sample goes to the left to node 2, otherwise it goes to the right to terminal node 3. Terminal nodes are stopping points and the majority of samples determine the classification of each terminal node. In terminal node 3, there is one control and 12 treated samples; the control is misclassified and the treated samples are classified correctly. Sensitivity is calculated as the ratio of the number of correctly classified treated samples to the total number of treated samples. Similarly, specificity is calculated as the ratio of the number of control samples to the total number of controls. (B) How the rules in the classification tree manifest themselves in the raw data. In this subset of six spectra from the study used to generate the tree in panel A, we can see how the classification tree uses the peak intensities to classify each sample.

PRODUCT & TECHNOLOGY REPORT

ples in our study. From here, we can apply practically all the tools used in analyzing gene expression data.

POST-PROCESSING I: FINDING SINGLE BIOMARKERS

Finding single genes or proteins responsible for differentiating disease vs. normal or treated vs. untreated is a natural first step in analyzing expression data (16). We use standard non-parametric statistical methods since we cannot assume that our peak intensity data conforms to a normal distribution and we oftentimes have a small sample size. These methods include the Mann-Whitney, the non-parametric equivalent of the Student's *t* test, and the Kruskal-Wallis, the non-parametric equivalent of ANOVA (analysis of variance), thus eliminating any assumption on the distribution of the peak intensity data (22). Essentially, these non-parametric tests sort the peak intensities and their corresponding ranks are used in the *P* value calculation. The *P* value results of these tests help to identify potential markers in conjunction with data visualization tools. Just as it is not practical to rely on views of microarray images once a large number is amassed in a single study, it is not always practical to view a large number of mass spectra. We instead rely on box-and-whisker plots and scatter plots both in log and linear scales as in Figure 2d.

These statistical tests simply give an indication of group mean differences, which may not always be helpful if we have a very large spread in the distributions of our data. Simply increasing the sample size can improve our *P* values while discrimination between groups may remain poor (20). This is especially important when attempting to use the biomarker in a diagnostic assay. Furthermore, we may not find single biomarkers with acceptable *P* values. At this point, we turn to other analytical methods that are both multivariate in nature and lend themselves towards developing a clinical assay.

POST-PROCESSING II: MULTI-DIMENSIONAL ANALYSIS

There are a number of different analysis tools that identify and use multivariate patterns in the expression profile data for the purposes of identifying groupings and/or classifying groupings. These methods can generally be lumped into one of two categories—unsupervised learning in the form of cluster analysis, or supervised learning in the form of classification methods. A number of these (only a fraction of existing literature is cited here) have been applied to gene expression data including clustering and visualization (4,8), self-organizing maps (11), and support vector machines (19). Although evaluating the plethora of methods to classify samples is an ongoing process, we have focused our attention on classification and regression tree-based methods (5,29), embodied in Biomarker Patterns™ Software. Similar to the way a doctor may make his or her diagnosis, “you are over 50, have hot flashes, and your estrogen level is low—therefore you are entering menopause,” the classification tree creates similar rules based on peak intensity such as, “peak intensity at 15153 Da < 5.169, and at 4400 Da > 12.283—therefore this sample is treated.” Figure 3 shows a sample tree and the spectral data underlying that tree. The algorithm in Biomark-

er Patterns Software examines each peak cluster present in the spectra and assesses its quality as a classifier. In addition, for each peak cluster that is determined to be a good classifier, Biomarker Patterns Software determines the intensity value that serves as a threshold above or below which a given classification is assigned.

A number of characteristics of classification trees make it an attractive tool for protein expression studies. The model is easy to interpret compared to “black-box” classifiers such as neural networks and nearest neighbor classifiers. The protein peaks used in the model are easily attainable by examination of the rules, and these rules are easily validated by examination of the spectra. This openness of the tree-based model is an attractive feature for researchers wanting both a diagnostic assay as well as potential therapeutic targets. In addition, classification trees can sift through all the input variables and select the subset to use in the tree. As such, it alleviates some of the burden of performing feature selection up front. However, it is always a challenge to perform feature selection in conjunction with any classifier, even in conjunction with classification and regression trees, and remains a topic of keen interest (3,18). In general, we strive to limit the number of input features to the number of samples in the dataset. Another feature of this software is that model building can be automated through the use of scripts written in virtually any language (C/C++, Perl, etc.).

Obviously, there will never be one method for classification that works best for all data sets, so we will continue to leverage the ongoing research in computational biology in the gene expression community as well as to investigate methods for better feature selection and protein expression data visualization. In addition, we continue our efforts towards better data management (2) by building database structures that handle the thousands of spectra generated in these clinical proteomics applications.

CONCLUSIONS

ProteinChip technology is a powerful tool for clinical proteomics that enables the rapid scanning of clinical samples to find diagnostic biomarkers. Like other genomic and proteomic technologies, this platform generates large amounts of data and the challenge is to sift through that data to find meaning. Informatic tools have been implemented to facilitate that process and additional informatics tools are being developed to further improve the workflow and mine additional, valuable biomarkers and biomarker patterns in SELDI-TOF-MS protein expression profiles.

ACKNOWLEDGMENTS

The authors would like to acknowledge the work of Leo Braginsky, Norton Kitagawa, Ed Gavin, Chris Apple, and Patrick Ho for their work on the preprocessing methods, as well as the other members of the software team.

REFERENCES

1. Austen, B.M., E.R. Frears, and H. Davies. 2000. The use of seldi proteinchip arrays to monitor production of Alzheimer's betaamyloid in transfected cells. *J. Pept. Sci.* 6:459-469.

2. Bassett, D.E., Jr., M.B. Eisen, and M.S. Boguski. 1999. Gene expression informatics—it's all in your mine. *Nat. Genet.* 21:51-55.
3. Ben-Dor, A., L. Bruhn, . Friedman, I. Nachman, M. Schummer, and Z. Yakhini. 2000. Tissue classification with gene expression profiles. *J. Comput. Biol.* 7:559-583.
4. Ben-Dor, A., R. Shamir, and Z. Yakhini. 1999. Clustering gene expression patterns. *J. Comput. Biol.* 6:281-297.
5. Breiman, L., J. Friedman, R. Olshen, and C. Stone. 1984. *Classification and Regression Trees*. Wadsworth, Pacific Grove.
6. Cai, H., Y. Wang, D. McCarthy, H. Wen, D.R. Borchelt, D.L. Price, and P.C. Wong. 2001. BACE1 is the major beta-secretase for generation of Abeta peptides by neurons. *Nat. Neurosci.* 4:233-234.
7. Chen, Y., E.R. Dougherty, and M.L. Bittner. 1997. Ratio-based decisions and the quantitative analysis of cDNA microarray images. *J. Biomed. Optics* 2:364-374.
8. Eisen, M.B., P.T. Spellman, P.O. Brown, and D. Botstein. 1998. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA* 95:14863-14868.
9. Fung, E.T., V. Thulasiraman, S.R. Weinberger, and E.A. Dalmasso. 2001. Protein biochips for differential profiling. *Curr. Opin. Biotechnol.* 12:65-69.
10. Fung, E.T., G.L. Wright, Jr., and E.A. Dalmasso. 2000. Proteomic strategies for biomarker identification: progress and challenges. *Curr. Opin. Mol. Ther.* 2:643-650.
11. Golub, T.R., D.K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J.P. Mesirov, H. Coller, M.L. Loh, J.R. Downing, M.A. Caligiuri, C.D. Bloomfield, and E.S. Lander. 1999. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* 286:531-537.
12. Griffin, T.J., S.P. Gygi, B. Rist, R. Aebersold, A. Loboda, A. Jilkine, W. Ens, and K.G. Standing. 2001. Quantitative proteomic analysis using a MALDI quadrupole time-of-flight mass spectrometer. *Anal. Chem.* 73:978-986.
13. Hampel, D.J., C. Sansome, M. Sha, S. Brodsky, W.E. Lawson, and M.S. Goligorsky. 2001. Toward proteomics in uroscopy: urinary protein profiles after radiocontrast medium administration. *J. Am. Soc. Nephrol.* 12:1026-1035.
14. Hegde, P., R. Qi, K. Abernathy, C. Gay, S. Dharap, R. Gaspard, J.E. Hughes, E. Snedrud, N. Lee, and J. Quackenbush. 2000. A concise guide to cDNA microarray analysis. *BioTechniques* 29:548-556.
15. Johnston-Wilson, N.L., C.M. Bouton, J. Pevsner, J.J. Breen, E.F. Torrey, and R.H. Yolken. 2001. Emerging technologies for large-scale screening of human tissues and fluids in the study of severe psychiatric disease. *Int. J. Neuropsychopharmacol.* 4:83-92.
16. Long, A.D., H.J. Mangalam, B.Y. Chan, L. Toller, G.W. Hatfield, and P. Baldi. 2001. Improved statistical inference from DNA microarray data using analysis of variance and a Bayesian statistical framework. *Analysis of global gene expression in Escherichia coli K12. J. Biol. Chem.* 276:19937-19944.
17. Merchant, M. and S.R. Weinberger. 2000. Recent advancements in surface-enhanced laser desorption/ionization time-of-flight mass spectrometry. *Electrophoresis* 21:1164-1177.
18. Model, F., P. Adorjan, A. Olek, and C. Piepenbrock. 2001. Feature selection for DNA methylation based cancer classification. *Bioinformatics* 17 Suppl 1:S157-164.
19. Moler, E.J., M.L. Chow, and I.S. Mian. 2000. Analysis of molecular profile data using generative and discriminative methods. *Physiol. Genomics* 4:109-126.
20. Pajak, T.F., G.M. Clark, D.J. Sargent, L.M. McShane, and M.E. Hammond. 2000. Statistical issues in tumor marker studies. *Arch. Pathol. Lab. Med.* 124:1011-1015.
21. Schadt, E.E., C. Li, C. Su, and W.H. Wong. 2000. Analyzing high-density oligonucleotide gene expression array data. *J. Cell. Biochem.* 80:192-202.
22. Sprinthal, R. 1997. *Basic Statistical Analysis*. Prentice-Hall, New Jersey.
23. von Eggeling, F., H. Davies, L. Lomas, W. Fiedler, K. Junker, U. Claussen, and G. Ernst. 2000. Tissue-specific microdissection coupled with ProteinChip array technologies: applications in cancer research. *BioTechniques* 29:1066-1070.
24. von Eggeling, F., K. Junker, W. Fiedler, V. Wollscheid, M. Durst, U. Claussen, and G. Ernst. 2001. Mass spectrometry meets chip technology: a new proteomic tool in cancer research? *Electrophoresis* 22:2898-2902.
25. Wang, S., D.L. Diamond, G.M. Hass, R. Sokoloff, and R.L. Vessella. 2001. Identification of prostate specific membrane antigen (PSMA) as the target of monoclonal antibody 107-1A4 by proteinchip array, surface-enhanced laser desorption/ionization (SELDI) technology. *Int. J. Cancer* 92:871-876.
26. Weinberger, S.R., S. Lin, P. Tornatore, D. King, and R. Orlando. 2001. Limited acid hydrolysis as a means of fragmenting proteins isolated upon ProteinChip® array surfaces. *Proteomics* 1:1172-1184.
27. Weinberger, S.R., T.S. Morris, and M. Pawlak. 2000. Recent trends in protein biochip technology. *Pharmacogenomics* 1:395-416.
28. Xiao, Z., B.L. Adam, L.H. Cazares, M.A. Clements, J.W. Davis, P.F. Schellhammer, E.A. Dalmasso, and G.L. Wright, Jr. 2001. Quantitation of serum prostate-specific membrane antigen by a novel protein biochip immunoassay discriminates benign from malignant prostate disease. *Cancer Res.* 61:6029-6033.
29. Zhang, H., C.Y. Yu, B. Singer, and M. Xiong. 2001. Recursive partitioning for tumor classification with gene expression microarray data. *Proc. Natl. Acad. Sci. USA* 98:6730-6735.
30. Zien, A., T. Aigner, R. Zimmer, and T. Lengauer. 2001. Centralization: a new method for the normalization of gene expression data. *Bioinformatics* 17 Suppl 1:S323-S331.

Address all correspondence to:

Eric Fung
 Ciphergen Biosystems
 6611 Dumbarton Circle
 Fremont, CA, 94555, USA
 e-mail: efung@ciphergen.com