Short Communication

Keith A. Baggerly
Jeffrey S. Morris
Jing Wang
David Gold
Lian-Chun Xiao
Kevin R. Coombes

Department of Biostatistics,
UT M.D. Anderson
Cancer Center,
Houston, TX, USA

A comprehensive approach to the analysis of matrix-assisted laser desorption/ionization-time of flight proteomics spectra from serum samples

For our analysis of the data from the First Annual Proteomics Data Mining Conference, we attempted to discriminate between 24 disease spectra (group A) and 17 normal spectra (group B). First, we processed the raw spectra by (i) correcting for additive sinusoidal noise (periodic on the time scale) affecting most spectra, (ii) correcting for the overall baseline level, (iii) normalizing, (iv) recombining fractions, and (v) using variable-width windows for data reduction. Also, we identified a set of polymeric peaks (at multiples of 180.6 Da) that is present in several normal spectra (B1–B8). After data processing, we found the intensities at the following mass to charge (*m/z*) values to be useful discriminators: 3077, 12 886 and 74 263. Using these values, we were able to achieve an overall classification accuracy of 38/41 (92.6%). Perfect classification could be achieved by adding two additional peaks, at 2476 and 6955. We identified these values by applying a genetic algorithm to a filtered list of *m/z* values using Mahalanobis distance between the group means as a fitness function.

Keywords: Cross validation / Data cleaning / Discrimination / Genetic algorithm / Mahalanobis distance
PRO 0522

Raw, not processed. The data set for the First Annual Proteomics Data Mining Conference consisted of both raw MALDI spectra and preprocessed lists of peak locations and heights [1]. We primarily used the raw spectra for our analysis, for two reasons. First, the reported intensities of the peaks are taken directly from the raw spectra without baseline correction or normalization. Further, the intensity levels of other spectra at this location are not recorded. Thus, the peak data effectively reduces to a binary matrix (present or absent), losing valuable intensity information. Second, several peaks that we could distinguish visually were not present in the processed data. This problem was more common for broad peaks at higher mass levels. Because several higher mass peaks correspond to proteins known to be useful in identifying disease conditions (*e.g.*, albumin at 66 kDa, immunoglobulin at 150 kDa), their omission is problematic [2].

Correspondence: Dr. Keith A. Baggerly, Department of Biostatistics, UT M.D. Anderson Cancer Center, 1515 Holcombe Blvd., Box 447, Houston, TX 77030, USA
E-mail: kabagg@mdanderson.org
Fax: +1-713-745-4940

Abbreviations: GA, genetic algorithm; **LDA**, linear discriminant analysis; **MD**, Mahalanobis distance

Sinusoidal noise removal. Visual inspection of the raw spectra revealed systematic distortions, particularly at the high *m/z* values: regular sinusoidal noise affected most of the spectra (Fig. 1). This noise was periodic on the time scale, not on the *m/z* scale. We applied a Fourier transform to several affected spectra, restricting the transform to regions where larger peaks were absent. The period of the noise (roughly 1760 clock ticks) was found to be nearly constant across different fractions and samples, but the phase appeared to be random. We suspect that this phenomenon is linked to the frequency of the alternating current in the power source, but cannot confirm this suspicion without more information. We are certain that it is not due to biology. Sinusoids of the appropriate frequency were fit to the tails of each spectrum, extended to the full spectrum length, and subtracted out. This processing is illustrated in Fig. 2.

Baseline subtraction. Visual inspection of the tails of the 20 fraction spectra from sample 1 showed variable baselines (Fig. 1). The minimum intensity ranged from 1900 to 2200 in flat regions (effectively zero), with peak heights rising about 60 units above this level. We considered several methods for baseline subtraction. Initially, we fitted a local median in a fixed window on the time scale. We chose a window size of 200 ticks, substantially wider than a peak at low intensity. Trial and error showed the

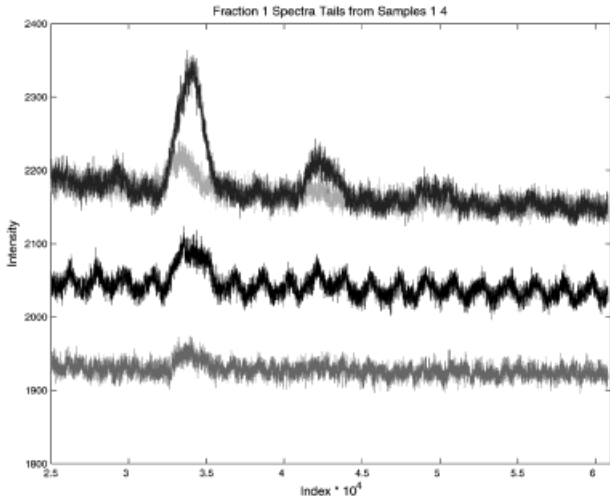

Figure 1. Tails of four spectra, on the time scale, showing fraction 1 from samples A1–A4. Sinusoidal noise is present in the middle spectrum (sample 4), and the spectra have different baselines.

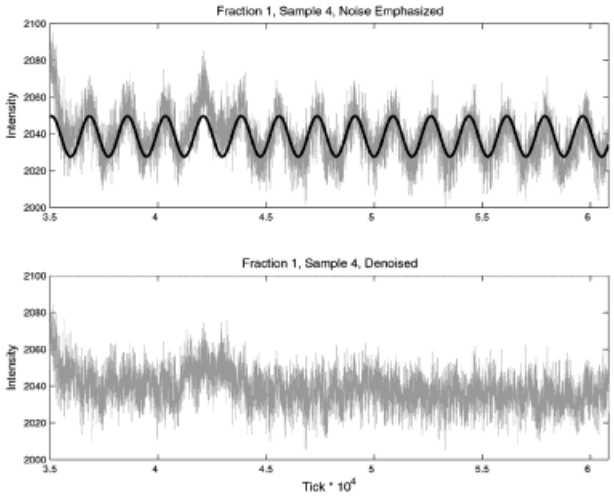

Figure 2. The last half of the spectrum for fraction 1, sample 4, before and after removal of sinusoidal noise.

results to be robust to changes in the window size down to 100 or up to 400. This method captured general trends well, especially at lower intensities where matrix elements were the main source of background noise. However, it tended to remove broad peaks at higher *m/z* values, which we wanted to retain. Subtracting the local median also made it difficult to assess the height of a peak. In order to locate zero accurately, we replaced the local median with the local minimum. In order to retain peaks in the high *m/z* range, we enforced monotonicity on the local minimum curve. Strict monotonicity was too stringent, because it introduced a bias at the right end of the

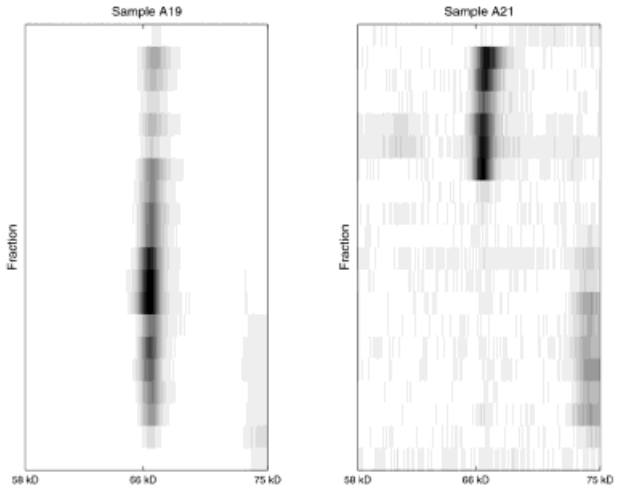

Figure 3. Artificial gels (or heat maps) from two diseased samples, centered around 66 000 Da, which is the mass of albumin. Twenty fractions are displayed vertically. The same protein migrates to different fractions in different samples. The horizontal axis is displayed in time units, and so it is nonlinear on the *m/z* scale.

plots. Our final decision was to subtract a "semimonotonic" baseline, defined as follows. First, we computed both the local and monotonic minimums. Next, we computed the median difference between the two minimums for the last 10 000 clock ticks in each spectrum. There appeared to be no signal in this region, so we wanted to introduce an offset level that would largely revert to the local minimum here. We took the minimum of twice this median difference and an absolute shift of 20 intensity units as the amount of "fuzz" allowed. Algorithmically, fuzz = min(20,2*median), semi-mono = min(mono + fuzz, local min). As an aside, the local minimum does not deal well with the sinusoidal noise, which is why we removed that first. It might be better to use a window whose width grew wider at higher intensities, but we have not pursued this idea.

Total ion current normalization. We normalized each baseline-corrected spectrum by dividing by the total ion current (the summed intensities over all time points). We also attempted to calibrate using a set of peaks, but had difficulty finding peaks that were stable across samples within a fraction, or across fractions within a sample.

Recombining fractions. After normalization, we produced images of the intensities from all spectra of a single sample as a function of fraction and *m/z*. These images showed a strong correlation between adjacent fractions. When we compared the plots for different samples, we found that some proteins (*e.g.*, albumin) migrated to different fractions in different samples (Fig. 3). Thus, comparing a single fraction across samples can be mislead-

ing: an apparently absent protein may be present in a different fraction. There appeared to be no pattern to this fractionation bias associated with disease status. Given this extra uncertainty, we elected to "undo" the fractionation by summing the normalized spectra from each sample and using this sum for further analysis. However, we retained information about peak locations from the individual fractions. The summed spectra contained evidence of a matrix bump at the low end, so they were baseline-corrected and normalized again.

The clock is visible in the spectra. Summing the corrected spectra uncovered an unexpected periodic phenomenon – a recurrent dip in intensity every $4096 = 2^{12}$ clock ticks. Smaller, more complicated periodicities occurred at other powers of 2. These periodicities differed from the sinusoidal noise discussed earlier. The sinusoidal noise was random in phase, and so largely canceled between spectra. Here, we were able to detect the new dip because of reinforcement across spectra. Further, this dip was uniformly present in all 41 averaged spectra. Because this phenomenon occurred at powers of 2, we strongly suspect that it is an artifact related to a computer chip inside the instrument recording the data.

An unknown polymer distinguishes eight samples. We computed two-sample *t*-statistics at each of the 60 831 *m/z* values in the recombined normalized spectra. Although this procedure did not yield any perfect classifiers, it did identify several peaks that distinguished normal samples B1–B8 from the rest. The peaks that were higher in these samples included several integer multiples of a mass of 180.6 Da, suggesting the presence of an unknown polymer with identical subunits (Fig. 4). We are uncertain as to the nature of this polymer, but it is not a protein since no amino acids or amino acid dimers have that molecular weight [3]. Glucose has a mass near 180, and polymerizes as starch, but in polymerizing it releases a water molecule and becomes too light. We may be detecting matrix elements or detergent.

Windowed dimension reduction. A visual inspection of the spectra in the regions of the most significant two-sample *t*-scores revealed that many of them occurred on slopes, rather than on peaks (data not shown). As a result, we decided to reduce the dimension of the space by windowing the data for each spectrum and choosing the maximum intensity within each window. This procedure partially corrects for the correlation in intensities at neighboring tick values. The window width in ticks varied smoothly (along a quartic polynomial) from 5 at low intensities to 500 at high intensities. Windowing reduced the dimension from 60 831 to roughly 2000. We further reduced the dimension by requiring each window to contain a peak. To define the existence of peaks, we used the

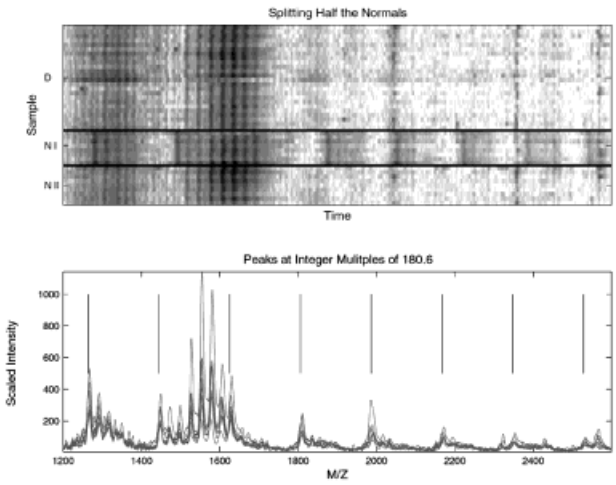

Figure 4. An unknown polymer occurs in eight samples. The top panel shows a heat map with samples displayed vertically and time displayed horizontally. The first eight normal samples (group N I) are clearly different. The bottom panel displays these eight spectra, highlighting peaks that occur at integer multiples of 180.6 Da.

processed data. A window was retained if a peak was detected in that range for at least eight of the samples. This step further reduced the dimension to 506.

Genetic algorithm with Mahalanobis distance. In order to focus on fold changes, we log transformed the 506 by 41 data matrix. We then searched for sets of $N = 1$, 2, 3, 4 and 5 peaks (rows) that achieved the best discrimination. A peak set was considered optimal if it maximized the Mahalanobis distance (MD), a multivariate generalization of the square of the two-sample *t*-test, between the two groups [4]. Mathematically, $MD = (\bar{x}_1 - \bar{x}_2)^T S_u^{-1} (\bar{x}_1 - \bar{x}_2)$, where S_u is the unbiased estimate of the covariance matrix.

This criterion finds the set of peaks giving the largest degree of separation between the centers of the two groups, assuming ellipsoidal cluster shapes. For $N = 1$ and 2 peaks, an exhaustive search was performed. Since combinatorics precluded exhaustive searching of the higher dimensional spaces, we used a genetic algorithm (GA) to find the best sets of $N = 3$, 4 and 5 peaks [5, 6]. For each N, we ran 50 replicate genetic algorithms using different randomly generated initial populations, each containing 200 sets of N peaks. The GA was allowed to evolve for 250 generations; each run of the GA converged. Convergence diagnostics for one representative run of the GA are shown in Fig. 5. For each set of peaks, we then obtained a separating hyperplane using Fisher's linear discriminant analysis (LDA). The best peaks found by our searches are listed in Table 1, which includes the

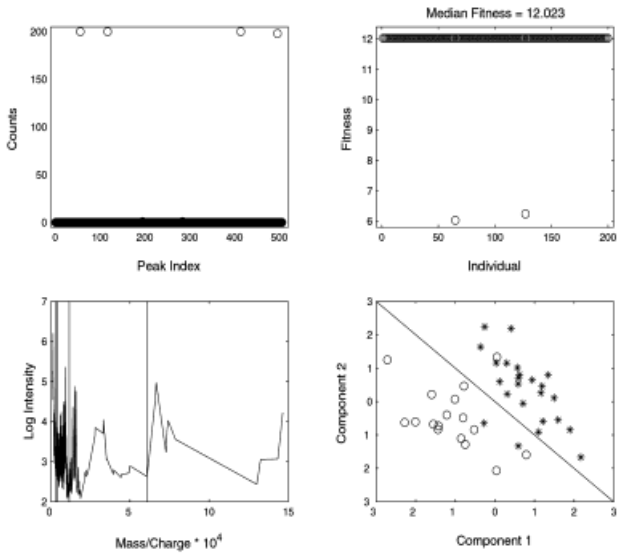

Figure 5. Final results of one run of the genetic algorithm looking for four peaks. The upper left panel shows how many times each of 506 peaks is included as part of an individual in the final population. The upper right panel shows the fitness (MD) of each of 200 individuals in the final population. The lower left panel shows the locations of the four most common peaks on the mass/charge scale. The lower right panel plots the two groups of samples using the first two principal components from the four peaks.

Mahalanobis distance between the two groups, the number of samples misclassified by the LDA, a theoretical estimate of the number of samples expected to be misclassified, the results of a leave-one-out cross-validation study, and an empirical *p*-value for the statistical significance of the observed Mahalanobis distance.

Expected number of misclassifications. Assuming that the two groups of samples have multivariate normal distributions with a common covariance structure, we can compute the expected probability of a misclassification as a simple function of the Mahalanobis distance: $\Phi\left(-\frac{1}{2}\sqrt{\text{MD}}\right)$, where Φ is the cumulative normal distribution [4]. The values in Table 1 were obtained by multiplying this probability by 41 and rounding to the nearest integer.

Leave-one-out cross validation. After completing a search for the best peak sets, we assessed the classification rule obtained from LDA using leave-one-out cross-validation. Specifically, using the selected peaks, we performed a separate LDA using only 40 of the 41 samples. We then used the results of the LDA to predict the status of the final sample. The number of times (out of 41) that the final sample was misclassified is included in Table 1.

Assessing statistical significance. Because the algorithm considers such a large number of peaks, it is important to check that the degree of discrimination given by our best set of peaks is greater in magnitude than expected due to random chance, *i.e.* in the case where there are in fact no differences between the groups' spectra. To assess this, we generated a "null distribution" of maximum MD for sets of peaks, which was simulated by finding the best sets of *N* peaks in samples randomly generated from a normal distribution with the same mean and covariance structure as our combined data set, pooling all 41 samples. The empirical *p*-value was computed as the proportion of simulated datasets having maximum MD greater than or equal to that found for our best sets of peaks.

Perfect classification with five peaks. We found sets of four or five peaks that give better classification rates than the sets listed in Table 1, albeit with smaller MD values. Performing LDA using the four peaks at 12 886, 17 318, 18 850 and 74 263 only misclassified one sample, with MD = 8.670. However, leave-one-out validation with these peaks misclassified five samples. Performing LDA by combining the optimum set of three peaks (at 3077, 12886 and 74263) with either the pair of peaks at 17318 and 61 000 (MD = 11.646) or the pair of peaks at 2476 and 6955 (MD = 11.161) gave a perfect classification. However, leave-one-out validation with these peaks misclassified three and two samples, respectively.

Table 1. Best sets of peaks for $N = 1, \ldots, 5$

Best peak set	Mahalanobis distance	Number misclassified	Expected number misclassified	Leave-one-out cross-validation	Empirical *p*-value
12 886	2.547	11	9	11	0.005
8840, 12 886	5.679	5	5	6	<0.01
3077, 12 886, 74 263	9.019	3	3	4	<0.01
5863, 8143, 8840, 12 886	12.585	3	2	3	<0.01
4125, 7000, 9010, 12 886, 74 263	23.108	1	1	1	<0.01

Nine peaks recur consistently. Finally, we looked at how often individual peaks occurred in the solutions found in multiple runs of the genetic algorithm. Some frequent peaks were adjacent to each other. These were checked visually to make sure that our binning had not inadvertently split a single peak across two bins; if it had we combined the results for the two bins. Our final results are summarized in Fig. 6. There were nine peaks that appeared more than twenty times, with bin centers at masses of: 3077, 4069, 5863, 6955, 8840, 12886, 17 318, 61 000 and 74 263. Using all nine peaks, we found MD = 19.948, with no misclassifications. Every solution set contained at least one of these nine peaks; most contained more than one. Moreover, visual inspection of the spectra confirmed that all nine peaks are present (Fig. 7).

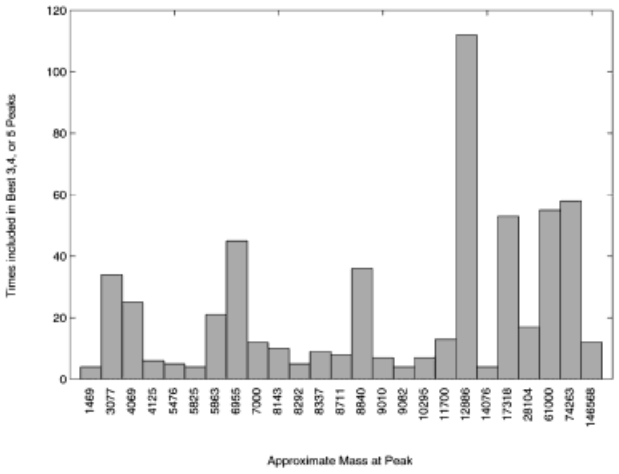

Figure 6. Number of times (out of 150 runs) that individual peaks were identified as part of an optimal solution by the genetic algorithm. The plot includes all peaks that were identified at least three times.

Best single peak. Some peaks are more important than others. In particular, the peak at 12 886 is the most interesting: it had the highest *t*-statistic, it appeared as a member of the best set of peaks for $N = 1, \ldots, 5$, and it appeared more times than any other peak.

Data preprocessing is extremely important. We have described our methods above; for a discussion of related issues, see [7]. There is still room for improvement in this area. There are complex interactions between baseline subtraction, normalization, noise estimation, and peak identification, so these steps should not be considered in isolation.

Dimension reduction is critical. The simulations we carried out with 1–5 peaks and 506 windows make us confident that the best peaks found in the actual data are unlikely to have arisen by chance. By contrast, we performed a handful of simulations with 2000 windows using five peaks and quickly found nearly perfect separation on random data (data not shown). We have no doubt that numerous methods could find perfect classifications using more "peaks" among the 60 831 *m/z* values across the 20 fractions in the raw spectra. However, we suspect that many of the "peaks" found that way would lack meaningful biological interpretation.

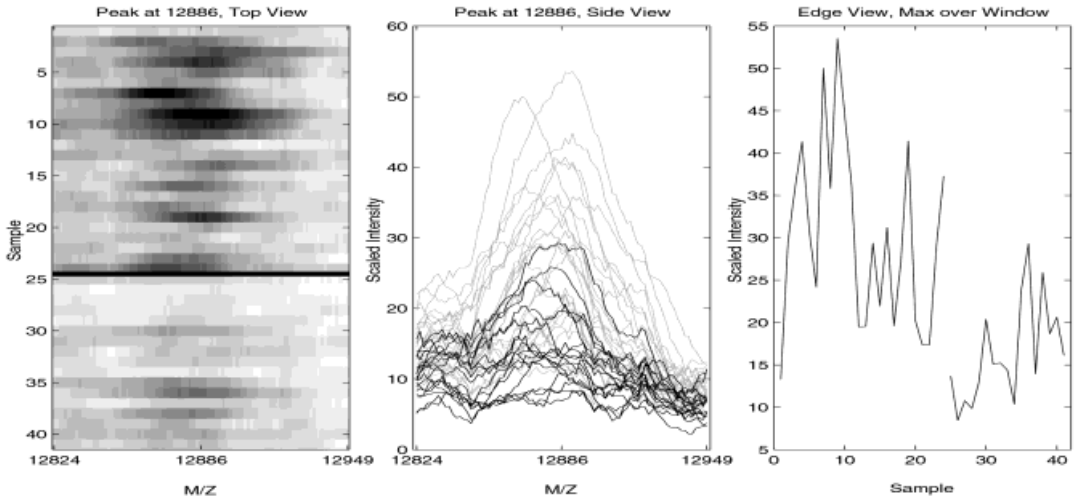

Figure 7. Final quality review of the peak at 12 886. The left panel is a heat map of the surrounding region with the vertical axis representing samples. Disease cases lie above the black bar; normal cases, below. The center panel displays all 41 spectra, shaded by disease status (light = disease, dark = no disease) centered near the peak. The right panel displays the height of the peak as a function of sample number. The first 24 samples are the disease cases.

GA+MD is good. The combination of a genetic algorithm with Mahalanobis distance is a simple but effective tool for finding sets of peaks that are different in the two groups. We considered more flexible discriminant functions, such as logistic regression or support vector machines, but these are slower. MD is a compromise between classification accuracy and computational efficiency. MD has recently been combined with a different directed random search method to find patterns that separate groups of samples in microarray data [8]. Genetic algorithms have also been combined with more elaborate clustering-based objective functions to study proteomics spectra [9].

We do not trust perfect classifications here. Although we were able to separate the data using a set of five peaks, the difference between this set of peaks and the set with the largest MD was minor. It is likely an accident that all the samples barely fell on the correct side of the linear separator between the groups. The results of the leave-one-out validation studies support this belief. There is no unique solution to the problem of finding sets of peaks that distinguish the samples in this data set. The objective function landscape being searched contains numerous maxima of similar magnitude. For this reason, it is important to run any randomized search process multiple times.

A compressed note. We note that zip compression is able to distinguish between the two groups, since the zip file when the groups are placed in subdirectories is smaller than the zip file with all spectra combined.

This work was partially supported by the Tobacco Settlement Funds as appropriated by the Texas State Legislature, and by a generous donation from the Michael and Betty Kadoorie Foundation to the Cancer Genomics Core Program. The authors also thank David Hawke and Ryuji Kobayashi for useful discussions during this work.

Received September 30, 2002

References

[1] Howard, B. A., Wang, M. Z., Campa, M. J., Corro, C. *et al.*, *Proteomics* 2003, *3*, 1720–1724.

[2] McPherson, R. A., in: Henry, J. B. (Ed.), *Clinical Diagnosis and Management by Laboratory Methods*, 19th edition, W. B. Saunders Company, Philadelphia 1996, pp. 237–252.

[3] Siuzdak, G. *Mass Spectrometry for Biotechnology*, Academic Press, New York 1996.

[4] Mardia, K. V., Kent, J. T., Bibby, J. M., *Multivariate Analysis*, Academic Press, Reading, MA 1979.

[5] Goldberg, D. E., *Genetic Algorithms in Search, Optimization, and Machine Learning*, Addison-Wesley, Reading, MA 1989.

[6] Holland, J. H., *Adaptation in Natural and Artificial Systems*, 3rd ed., MIT Press, Cambridge, MA 1994.

[7] Fung, E. T., Enderwick, C., *Biotechniques* 2002, *32*, *Suppl*. 34–41.

[8] Chilingaryan, A., Gevorgyan, N., Vardanyan, A., Jones, D., Szabo, A., *Math. Biosci.* 2002, *176*, 59–72.

[9] Petricoin, E. F. III, Ardekani, A. M., Hitt, B. A., Levine, P. J. *et al.*, *Lancet* 2002, *359*, 572–577.