Disease Markers 19 (2003,2004) 169–183
IOS Press

The use of urine proteomic and metabonomic patterns for the diagnosis of interstitial cystitis and bacterial cystitis

Que N. Van[a], John R. Klose[a], David A. Lucas[a], DaRue A. Prieto[a], Brian Luke[b], Jack Collins[b], Stanley K. Burt[b], Gwendolyn N. Chmurny[a], Haleem J. Issaq[a], Thomas P. Conrads[a], Timothy D. Veenstra[a] and Susan K. Keay[c,d,*]
[a]*Laboratory of Proteomics and Analytical Technologies, SAIC-Frederick, Inc., NCI Frederick, Frederick, MD, USA*
[b]*Advanced Biomedical Computer Center, SAIC-Frederick Inc., NCI-Frederick, Frederick, MD, USA*
[c]*Division of Infectious Diseases, Department of Medicine, University of Maryland School of Medicine, Baltimore, MD 21201, USA*
[d]*Research Service, VA Maryland Health Care System, Baltimore, MD 21201, USA*

Abstract. The advent of systems biology approaches that have stemmed from the sequencing of the human genome has led to the search for new methods to diagnose diseases. While much effort has been focused on the identification of disease-specific biomarkers, recent efforts are underway toward the use of proteomic and metabonomic patterns to indicate disease. We have developed and contrasted the use of both proteomic and metabonomic patterns in urine for the detection of interstitial cystitis (IC). The methodology relies on advanced bioinformatics to scrutinize information contained within mass spectrometry (MS) and high-resolution proton nuclear magnetic resonance (^{1}H-NMR) spectral patterns to distinguish IC-affected from non-affected individuals as well as those suffering from bacterial cystitis (BC). We have applied a novel pattern recognition tool that employs an unsupervised system (self-organizing-type cluster mapping) as a fitness test for a supervised system (a genetic algorithm). With this approach, a training set comprised of mass spectra and ^{1}H-NMR spectra from urine derived from either unaffected individuals or patients with IC is employed so that the most fit combination of relative, normalized intensity features defined at precise *m/z* or chemical shift values plotted in n-space can reliably distinguish the cohorts used in training. Using this bioinformatic approach, we were able to discriminate spectral patterns associated with IC-affected, BC-affected, and unaffected patients with a success rate of approximately 84%.

1. Introduction

With the rapid development of methods in the fields of genomics (DNA), transcriptomics (mRNA), proteomics (proteins), and metabonomics (low molecular weight metabolites) there is general enthusiasm towards revolutions in systems biology that will lead to more advanced approaches to diagnostics and therapeutics. Much of the effort in these areas focuses on comparing thousands of species between unaffected and diseased individuals with the hope that one, or a few, key differences in the two states may be identified. While ideally these differences would be recognized in readily obtainable biofluids such as urine, plasma, or serum, the inter-person variability of these samples makes the identification of unique, disease-reflective differences quite challenging. While unique biomarkers, such as HCG for pregnancy, are extremely effective, others such as Cancer Antigen 125 and prostate specific antigen possess poor positive-predictive value – particularly for early disease stage diagnosis.

*Corresponding authot: Dr. Susan Keay, VA Medical Center, Room 3B-184, 10 N. Greene Street, Baltimore, MD 21201, USA. TeL.: +1 410 605 7000 ext. 6450; Fax: +1 410 605 7837; E-mail: skeay@umaryland.edu.

ISSN 0278-0240/03,04/$17.00

Petricoin et al. have recently demonstrated that low molecular weight serum proteomic patterns from surface-enhanced laser desorption ionization time-of-flight mass spectral (SELDI TOF-MS) data can distinguish neoplastic from non-neoplastic disease within the ovary [16]. A key aspect to their study was the application of a high-order self-organizing cluster analysis approach based on a genetic algorithm that was "trained" on SELDI-TOF MS spectra from serum derived from either healthy women or women with ovarian cancer. The "trained" algorithm was applied to a masked set of samples and resulted in a sensitivity of 100%, a specificity of 95% and a positive-predictive value of ovarian cancer of 94% [16]. The success of the use of proteomic patterns for the diagnosis of stage I ovarian cancer suggests that patterns generated from other biomolecules within biofluids may also provide a useful indicator of the early onset of a particular disease state.

Since proteomic patterns of serum acquired using SELDI TOF-MS can be diagnostic of a particular disease state, it follows that spectral patterns of biofluids acquired using other types of analytical techniques may also be useful diagnostic tools. Nuclear magnetic resonance (NMR) spectroscopic analysis of bulk biofluids such as urine or plasma (e.g. metabonomics) has been utilized as a means to measure time-related biochemical responses resulting from physiological, pathological, or interventional genetic events [12–14]. High-field proton (^{1}H) NMR spectra of biofluids typically contain several hundred resolvable lines, potentially providing structural and quantitative information on hundreds of compounds in a single, nondestructive analysis that takes only a few minutes. The resulting spectrum provides a profile of the metabolic status of the organism. Recently, Brindle et al. showed the capability of discriminating serum samples acquired from patients with coronary heart disase from those with angiographically normal coronary arteries by analyzing the ^{1}H-NMR spectra of each sample using a supervised partial least squares discriminant algorithm [1]. This non-invasive method was shown to have a specificity of $>$90%.

We studied the effectiveness of analyzing MS and ^{1}H-NMR data using a genetic algorithm combined with a self-organizing cluster analysis to correctly discriminate urine samples from individuals suffering from interstitial cystitis (IC) from those of healthy individuals. IC is a debilitating chronic bladder disease of unknown etiology that affects an estimated 750,000 women in the United States, with one-tenth as many men also diagnosed with this disease [2,15,17,18]. IC is currently diagnosed only by symptomatic criteria (urinary frequency plus pain and/or urgency) in the absence of specific identifiable causes, combined with cystoscopic findings (including petechial hemorrhages called "glomerulations" in approximately 90% of patients, or ulcers that extend into the lamina propria in approximately 10%) [3,6,20]. None of these symptoms, however, are specific for IC, and the specificity of glomerulations for this disorder has also been called into question [21], making it currently difficult to establish the diagnosis of IC in a particular patient. Several urine biomarkers have been associated with IC that ultimately may prove to be useful for the noninvasive diagnosis of this disorder, including an antiproliferative factor (APF) that inhibits the proliferation of normal primary human bladder epithelial cells *in vitro* [4,7,9], heparin-binding epidermal growth factor-like growth factor, and epidermal growth factor [4,8,10]. Additional noninvasive diagnostic criteria based on urine or serum markers would be useful for establishing the diagnosis of IC as well as for understanding the pathogenesis of this disorder. In addition, to determine the specificity of findings related to IC specimens, we generated proteomic and NMR spectral patterns of urine samples from people suffering from acute bacterial cystitis (BC). In the following we describe the use of MS and ^{1}H NMR spectra of urine obtained from patients with IC, patients with BC, and unaffected controls to identify those patients with interstitial cystitis.

2. Materials and methods

2.1. Patients

All 40 female and 10 male IC patients had previously undergone cystoscopy and fulfilled the NIDDK diagnostic criteria for IC [3]. In addition, 30 females were identified as having acute bacterial cystitis (diagnosed by the presence of bacteriuria with $>10^3$ of a single type of bacteria per milliliter of urine, plus pyuria, in combination with appropriate symptoms). Asymptomatic controls included individuals that were age- ($\pm$ 5 years), race- and sex-matched to the IC patients (i.e. 40 females and 10 males). All participants were at least 18 years old and were enrolled in accordance with guidelines of the Institutional Review Board of the University of Maryland School of Medicine.

2.2. Urine specimens

Urine was collected by the clean catch method in which each IC patient, bacterial cystitis patient, or control wiped the labial area with 10% povidone iodine solution and then collected midstream urine into a sterile container. Specimens were initially kept at 4°C, then transported to the laboratory where cellular debris was removed by low speed centrifugation at 4°C. Urine samples were adjusted to pH 7.2 (using 10 N HCl or 10 N NaOH) and 300 mOsm (using 1 M NaCl or ddH_2O), and filtered through a 0.2 μm pore size filter (Gelman Sciences, Ann Arbor, MI). Each specimen was aliquoted under sterile conditions and stored at –80°C.

2.3. ProteinChip array sample preparation

WCX2 ProteinChip arrays were loaded into a 96-well bioprocessor (Ciphergen Biosystems Inc., Palo Alto, CA) and activated with 10 mM HCL. Arrays were washed with HPLC-grade water and pre-equilibrated with binding buffer (50mM sodium acetate, pH 4.5). One hundred μL of urine (diluted 1:1 in binding buffer) was added in duplicate to the WCX-2 ProteinChip array surface and incubated for 3 hours at ambient temperature with gentle agitation. The ProteinChip arrays were washed three times with 100 μL of binding buffer, followed by a final wash with 100 μL of HPLC-grade water. ProteinChip arrays were removed from the bioprocessor and air-dried. One μL of 20% α-cyano-4-hydroxycinnamic acid solution in 50% acetonitrile, 0.5% trifluroacetic acid was added to each WCX-2 ProteinChip array bait surface.

2.4. PBS-II TOF MS analysis

ProteinChipTM arrays were analyzed by a Protein Biological System II time-of-flight mass spectrometer (PBS-II, Ciphergen Biosystems Inc.) and mass spectra were recorded using the following settings: laser intensity 185, detector sensitivity 8, *m/z* range 0–20,000, 130 shots per sample. Data were collected using the Ciphergen ProteinChip software version 3.0. The PBS-II TOF MS was externally calibrated using the "All-In-One" peptide mass standard (Ciphergen Biosystems, Inc.).

2.5. Proteomic pattern analysis

Proteomic pattern analysis was performed by exporting the raw data files generated from the PBS-II into tab-delimited files possessing approximately 15,000 data points. The mass spectra were randomly segregated into equal groups for training, and blind testing. The models were built on the training set using ProteomeQuestTM (Correlogic Systems Inc., Bethesda, MD) and tested using blinded sample sets. The Proteome QuestTM software itself implements a pattern discovery algorithm combining elements from genetic algorithms [5] and self-organizing adaptive pattern recognition systems [11]. Genetic algorithms organize and analyze complex data sets as if they were information comprised of individual elements that can be manipulated through a computer-driven analog of a natural selection process. Self-organizing systems cluster data patterns into similar groups. Adaptive systems recognize novel events and track rare instances. The genetic algorithm component of analysis begins with the random generation of a population of 1500 subsets of combinations of features in the urine mass spectra. This number was chosen based on adequate coverage of the data, with a heuristic that no value can be duplicated within each of the 1500 feature subsets. Each feature subset in the population specifies the identities of the exact *m/z* values in each urine mass spectrum but not their relative amplitude. The number of features in the subset ranges from 5 to 20. For this study, MS data was normalized by linearly scaling each *m/z* value, V, within any randomly generated pattern subset between the largest and the smallest values within that subset so that $0 \leqslant \mathrm{NV} \leqslant 1$. In this way, differences in spectral quality that may emanate from biases such as in ProteinChip variance and not from the inherent disease process itself can be minimized. The spectra are normalized according to the following formula:

$$\mathrm{NV} = (\mathrm{V} - \mathrm{Min})/(\mathrm{Max} - \mathrm{Min})$$

Where NV is the normalized *m/z* value, V is the intensity value for the specific randomly chosen *m/z* bin, Min is the intensity of the smallest intensity value of any of the *m/z* bins within the randomly selected feature set and Max is the maximum intensity of the *m/z* bin within the randomly selected feature set. This equation linearly normalizes the peak intensities in the feature set so as to fall within the range of 0 to 1. Prior to analysis, the data is randomly divided into training and testing data sets. The training data set is further divided into and labeled as diseased or unaffected based upon known clinical diagnosis.

2.6. NMR data acquisition

Before ^{1}H-NMR data acquisition, each urine sample was equilibrated to ambient temperature. A D_2O stock solution containing 0.21% (w/v) sodium 4,4-dimethyl-4-silapentanoate- 2,2,3,3-d_4 (TSP) was prepared by dissolving 10.5 mg of TSP in 5 mL of D_2O. Thirty-three μL of this solution was added to each 325 μL urine sample, which was then vortexed and transferred to a 5 mm Shigemi NMR tube with 15 mm susceptibility matched plungers. ^{1}H-NMR spectra were acquired of urine samples obtained from 47 control, 50 IC-affected, and 30 BC-affected individuals.

NMR spectra were acquired on a 500 MHz Varian INOVA Spectrometer equipped with a Nalorac indirect gradient HCNP probe and using the B1-insensitive WET water suppression pulse sequence as described by Smallcombe et al. [19]. The WET selective pulses were a 6 ms Gaussian at 8.1 dB. The gradients were 2 ms in length with levels of 24000, 12000, 6000, and 3000, respectively, each followed by a 2 ms delay. The spectra were collected at 27°C with a spectral width of 6500 Hz, 5 s acquisition, 5 s equilibrium delay, 32 transients preceded by 1 steady state transients, and a 45° acquisition pulse (4.5 μs at 56 dB). The transmitter was set on the water resonance at –175.5 Hz and was not changed from one sample to the next. The probe was retuned for each sample.

2.7. Metabonomic pattern analysis

The 32499 complex points from each ^{1}H-NMR spectrum was zero-filled to the next power of two, 32768 complex points, and transformed with 0.5 Hz exponential line broadening. Each spectrum was phased, referenced to TSP and drift corrected. The Varian's binning package, provided by Dr. Bruce Adams, was used to convert each spectrum into 531 bins starting from 0.16 ppm to 10.80 ppm with widths of 0.02 ppm. The integration value for bins in the regions between 4.60–4.88 ppm and 5.52–6.04 ppm were set to zero to remove contributions from the residual water and urea peak respectively. The data was normalized by scaling the sum of the 531 bins for each spectrum to a value of 50,000.

2.7.1. Identification of spectral outliers

The 127 binned NMR spectra were normalized so that each has a binned intensity that sums to 50,000. Prior to classification, the data were analyzed for the presence of any strikingly different spectra (i.e. an "outlier") from all others within the cohort of 127 spectra. Outliers were identified using either principal component analysis (PCA) or a Sammon Map [1]. A Sammon Map is a projection of a high-dimensional set of data onto a lower-dimensional space such that the distance between all pairs of data points is preserved to the greatest extent. If $D_{i,j}$ is the calculated distance between a pair of cohorts and $d_{i,j}$ is the distance in the lower-dimensional space, the Sammon mapping tries to minimize the following metric.

$$\sum_{i=2,N}\sum_{j<i}\{(D_{i,j}-d_{i,j})^2/D_{i,j}\}/\sum_{i=2,N}\sum_{j<i}D_{i,j}$$

In this study, each of the 127 NMR spectra was projected onto a 2-dimensional plot by randomly placing each cohort in a plane and then performing 400 Newton-Raphson optimizations of each coordinate to minimize the above expression. This procedure is repeated 400 times, and the 2-dimensional mapping that yields the lowest metric is used.

Instead of using the difference-squared as a measure of the disagreement between two cohorts in a given bin, the agreement between their overall profiles can be used as a measure of their similarity. By comparing the intensities in each bin for two cohorts, the sum of the minimum intensities represents the overlap in their profiles. This overlap is equivalent to the summed intensity (50,000) minus one-half of the Manhattan distance between them. The Manhattan distance (L1-norm) is simply the sum of the absolute difference in intensities. The percent similarity is then 100.0 times the overlap divided by 50,000. The similarity matrix is then used in a corresponding K-Most Similar Neighbor analysis using the same predictive procedure as above.

These procedures that use the overall NMR profiles are unbiased, but may be strongly affected by dietary or other random factors. In addition, though they use the magnitude of the differences in the profiles, they do not determine where the profiles are significantly correlated to the Class of the cohort and therefore yield no information about a possible metabolite that may be useful in the classification. In an attempt to find significant bins, a feature selection method is used to select a small number of bins and only the intensities in these bins are used to construct a distance matrix. This matrix is then used in a K-Nearest Neighbor algorithm that is slightly different from that described above.

2.7.2. Distance-dependent K-nearest neighbors analysis

A modified evolution programming (EP) method was used to identify features that can classify the NMR

spectra as being obtained from urine acquired from IC-affected, BC-affected, or healthy patients. An EP was selected because it allows the parent population to maintain diverse solutions from one generation to the next, thereby allowing the final population from this method to be used as a good starting population for a new search if new samples are added to the analysis. With the other three methods, this new search would have to start from scratch since the population is homogeneous. The same argument applies if, upon analysis of the features selected, it is found that one or more of the features have no biological basis. This feature can be randomly changed to another feature in any members of the final EP population that contain it and the search can continue, while the other methods would again have to start from scratch since the new search would be limited to a one-dimensional search of the replaced feature.

In the EP feature selection method used here, a population of N genetic vectors of length L ($L = 4$ or 8 here) was randomly generated. Each set of L features was then used in the modified KNN procedure described below to generate a cost function that measures the degree to which the cohorts are incorrectly classified. In each generation, each parent generates a new genetic vector by randomly replacing one or two of the features in the parent's genetic vector with new features. One of the features is required to be replaced while the second is probabilistically replaced. In the results presented here, the probability of a second replacement is 50% in the first generation and linearly decreases to 1% in the last generation. Before the cost of this offspring is determined, its genetic vector is compared with all genetic vectors in the parent population and all vectors of offspring that it has produced so far. If it is found to be the same as any existing solution, this offspring is destroyed and the same parent is used to generate a new offspring. This *uniqueness operator* represents one of many possible *maturation operators* that can be used with the EP method and guarantees that the parent population will be diverse from generation to generation. A generation is complete once each parent has generated a unique offspring and the offspring's cost has been determined. At the end of each generation a $(\mu + \lambda)$ deterministic selection procedure is used to select the parents for the next generation. This process means that the parent and offspring populations are combined to form a population with $2 \times N$ solutions, and the N solutions with the lowest cost become parents in the next generation. This process is continued for M generations, at which time the search stops and the 50 feature sets with the lowest cost are reported.

When each set of L features is examined, the intensities in these bins maps each spectrum onto a point in L-dimensional space. A Manhattan and Euclidean distance metric was used to determine the distance between each cohort-pair and construct a distance matrix. In the KNN procedure outlined above, the K nearest samples to a given sample are used to predict its class (i.e. IC, BC, or healthy). In the case of 4-nearest neighbors, the probability that this cohort belongs to a certain class can be 0, 25, 50, 75, or 100%, depending upon the classification of the four closest samples, and independent of the distance they are from the given cohort. In this analysis, the distances to the four closest neighbors affect the prediction.

The classification method used in this analysis is actually a Distance-Dependent KNN. In a DD-KNN classification if one of the four neighbors is significantly closer to the given sample than the rest, its classification influences the prediction more than the others. Similarly, if the given sample is far away from any of its neighbors, its classification is less certain. If one of the nearest neighbors is Cohort-i with a classification of Cl(i), the unnormalized probability that this cohort belongs to this Class, p[Cl(i)], is a determined by a monotonically decreasing function of the distance between the given cohort and Cohort-i. If d(i) is the distance from a given sample to Neighbor-i, the unnormalized probability of being in the same class is

$$p[Cl(i)] = a/d(i)$$

This unnormalized probability is truncated at a large value if d(i) is sufficiently small. This function increases the probability that the cohort has the same class as a neighbor if the neighbor is close, but does not take care of the case where the cohort is far away from all neighbors. To handle this case, a fourth classification called *Unknown* is added and the unnormalized probability that the cohort belongs to this class relative to each neighbor is given by the expressions

$$\begin{aligned} &= Pu \quad \text{if } p[Cl(i)] \leqslant (1 - 2Pu) \\ p[\text{Unknown}] &= (1 - p[Cl(i)])/2 \text{ if } (1 - 2Pu) \\ &< p[Cl(i)] \leqslant 1.0 \\ &= 0 \ \text{ if } p[Cl(i)] > 1 \end{aligned}$$

In the results presented here, Pu is set to 0.1, meaning that the unnormalized probability of the cohort belonging to the Unknown-Class for a given neighbor is 0.1 if the probability that it is the same class as this neighbor is 0.8 or less; it decreases from 0.1 to 0.0 as the probability of being in the same class as the neighbor

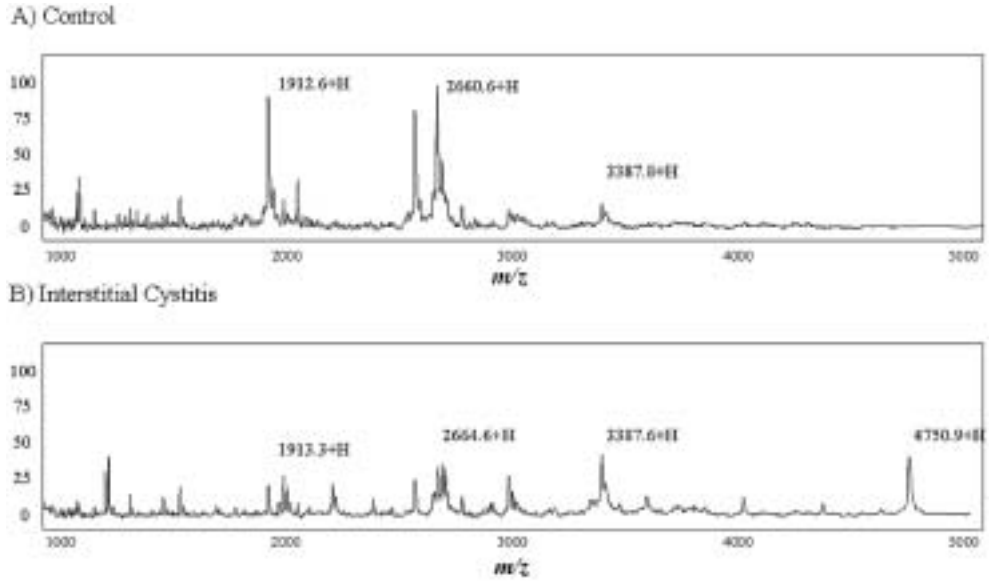

Fig. 1. Comparison of SELDI-TOF MS profiles of urine samples obtained from (A) normal controls and (B) interstitial cystitis patients.

increases to 1.0; and is 0.0 if the unnormalized probability of belonging to the neighbor's Class is greater than 1.0.

After adding the contributions to the unnormalized probabilities of the three p[Cl(i)]'s and p[Unknown] from each of the four nearest neighbors, they can be normalized by division by their sum. If the given cohort belongs to Class-I, the error in the characterization of this cohort is simply (1-P(I)), where P(I) is the predicted, normalized probability that it is in this Class based upon its four neighbors. By summing this error for all cohorts, the Cost of this set of features is obtained.

The last requirement is to set a value for α in the expression for p[Cl(i)] above. This constant is determined by the user-supplied value of HALF which controls the value of $d_h(i)$ such that

$$p[Cl(i)] = 1/2 \text{ when } d(i) = d_h(i)$$

Since the magnitudes of the intensities changes for different bins, $d_h(i)$ is set to HALF times the theoretical maximum distance (TMD) between cohorts for a given set of bins (features). TMD is determined by using the difference between the maximum and minimum intensities in each of the selected bins. α is then determined from the expression

$$a = 1/2 \times HALF \times TMD$$

In the results presented here, both scaled and unscaled differences in intensities are used to calculate either the Manhattan or Euclidean distance between two samples (and the TMD). The scaled difference is the absolute difference divided by the average of the intensities and this option produces a relative change instead of an absolute change. In addition, HALF is set to 0.1, 0.15 and 0.2 in different runs to study the effect of increasing α.

The next section therefore presents the results of 24 different classification runs. The EP method searches for the optimum set of either four or eight features and the Cost of each set is determined from one of 12 Distance-Dependent KNN examinations (two possible differences, two distance metrics, and three values of α).

3. Results

Urine samples were collected from both healthy individuals as well as those had previously undergone cystoscopy and fulfilled the National Institute Diabetes and Digestive and Kidney Diseases (NIDDK) diagnostic criteria for IC [3]. Prior to ^{1}H-NMR analysis the samples were adjusted to pH 7.2 and 300 mOsm and filtered through a 0.2 μm pore filter. The MS and ^{1}H-NMR spectra of a selection of urine samples from both healthy and IC-affected individuals are shown in Figs 1 and 2, respectively. A comparison of the spectra showed that while there is variability between those acquired from the two sample sets, there also exists vari-

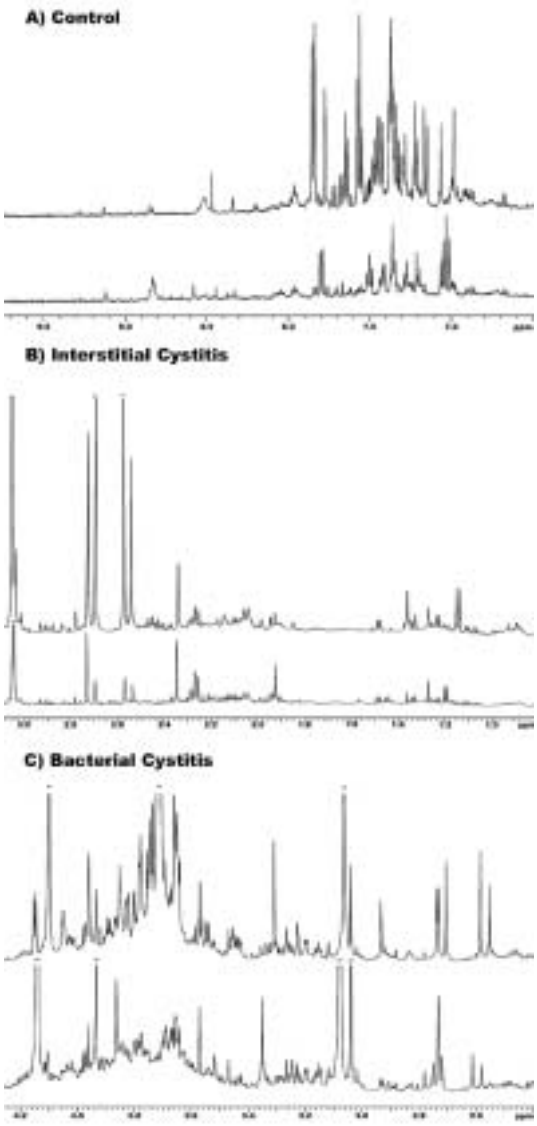

Fig. 2. Comparison of ^{1}H-NMR spectra of urine samples obtained from (A) normal controls, (B) interstitial cystitis patients, and (C) bacterial cystitis patients.

ability within a single sample set. This inherent variability makes it difficult to visually identify signals that are consistently unique to either the control or disease samples, requiring the application of analytical methods that utilize bioinformatic algorithms to distinguish patterns between these three groups.

3.1. Diagnosis of urine samples by proteomic pattern analysis

Proteomic pattern analysis was performed by exporting the raw data file generated from the PBS-II TOF-MS. The training set consisted of MS spectral data accumulated from urine samples obtained from 14 asymptomatic controls and 29 patients with IC. The models were built on the training set using ProteomeQuestTM and blind testing was performed with 16 control and 21 IC urine samples. The *m/z* (their intensities) that were found to be classifiers used to distinguish urine from a patient with IC from that of an unaffected individual are based on actual values from the raw MS spectra. A total of eight different *m/z* classifiers (*m/z* 2980.07, 3939.11, 4003.20, 4391.23, 5386.83, 9769.87, 10090.55, and 18893.23) were required to correctly segregate the urine samples obtained from healthy vs. IC-affected individuals. Blinded testing of this model generated from the training set resulted in 100% sensitivity and specificity for the diagnosis of the spectra obtained of the urine samples from the 16 control and 21 IC-affected individuals. Obviously this level of sensitivity and specificity only applies to this limited sample set; whether this diagnostic accuracy can be achieved over a larger cohort would need to be determined by conducting a larger trial.

3.2. Classification of source of urine samples by metabonomics pattern analysis

The encouraging results obtained in the proteomic profiling of urine samples from healthy and IC-affected individuals prompted us to investigate whether similar diagnostic capabilities could be obtained using ^{1}H-NMR. In addition, we sought to investigate whether the ^{1}H-NMR spectra acquired from urine samples obtained from more than two groups of patients with specific disorders could be segregated. For this investigation, we tested alternative bioinformatic algorithms with the goal of distinguishing IC patients not only from healthy controls, but also from patients with bacterial cystitis (BC). While there are much more cost effective methods to diagnose BC rather than the use of NMR, nonetheless it would be crucial to develop methods that could effectively diagnose IC with a high positive predictive value. Therefore, it was important to determine if NMR could be used to segregate urine samples obtained from three distinct conditions with a low rate of false positive identiification.

3.2.1. Search for outlying spectra

A plot of the first versus the second Principal Component (PC) of all of the ^{1}H-NMR spectra acquired was constructed in order to identify outlier spectra that may arise from either errors in sample collection, process-

ing, or data acquisition, as shown in Fig. 3. A plot of the first few PCs is not guaranteed to reveal outliers, but is sufficient for this dataset because the first PC accounts for approximately 50.3% of the total variation in the data within this particular spectrum and an analysis of this component shows that the coefficient for bin/feature 332 is -0.9936. Since this component is virtually composed of a single feature, its coefficient is negative, and one sample spectrum has a large negative value relative to the rest, this cohort has an intensity that is many times larger than for any other cohort. Because the difference is concentrated in a single feature, this feature will have a large variance and it will be one of the first few PCs. If, on the other hand, the difference between this cohort and the rest was spread out across all of the features, it may not appear as an outlier in this type of plot.

To confirm the presence of the outlier spectrum determined by PCA, a two-dimensional Sammon Map of the set of 127 samples was generated (data not shown). This map attempts to conserve the inter-cohort distances to the largest possible extent, and it also confirms the presence of a single outlier recognized by PCA. This outlier would appear if the differences between it and the other cohorts are uniformly distributed across all features or, as in this case, it is concentrated in a single feature. Since this outlier can adversely affect subsequent classification studies it is removed from consideration. The outlier identified by the Sammon Map corresponds to the exact outlier identified by PC analysis (described above). Therefore, the classification is only performed on the remaining 126 spectra (46 controls, 50 IC patients, and 30 BC patients).

3.2.2. Classification by distance-dependent K nearest neighbors

The Euclidean distance matrix used to construct the Sammon Map was also used in a standard KNN study, as shown in Table 1(A). This table shows that if only the nearest neighbor is used to predict the classification of each cohort ($K = 1$), the set of 126 cohorts are correctly classified 76.19% of the time. Because 1-Nearest Neighbor is deterministic instead of probabilistic, this number shows that 96 cohorts are correctly classified and 30 are not. The 46 control cohorts are correctly classified in 34 cases and misclassified in 12; 43 of the 50 IC patients are correctly classified and seven are not; and 19 of the BC patients are correctly classified while 11 are not. This breakdown is not possible when the number of neighbors exceeds one since for 2-Nearest Neighbors a given cohort can be 100% correctly classified, 50% correctly classified, or completely misclassified. Table 1 shows that the quality of the classification decreases as the number of neighbors increase.

Table 1
Probabilities of correct classification for the 126 cohorts when the (A) Euclidean distance and (B) Similarity (Manhattan distance) between their overall ^{1}H-NMR profiles is used in a K-Nearest Neighbor study

A.				
	$K = 1$	$K = 2$	$K = 3$	$K = 4$
Overall	76.19%	72.62%	67.20%	63.49%
Normal controls	73.91%	71.74%	64.49%	61.96%
IC patients	86.00%	79.00%	78.00%	75.00%
BC patients	63.33%	63.33%	53.33%	46.67%
B.				
	$K = 1$	$K = 2$	$K = 3$	$K = 4$
Overall	86.51%	81.35%	77.78%	71.63%
Normal controls	82.61%	78.26%	77.54%	73.37%
IC patients	94.00%	89.00%	87.33%	79.50%
BC patients	80.00%	73.33%	62.22%	55.83%

When the Manhattan distance matrix is constructed to determine the similarity between the NMR profiles and is then used in a KNN (or K-Most_Similar Neighbors) classification study, the results in Table 1 (B) are obtained. A comparison with Table 1(A) shows that this distance metric yields consistently better results. For 1-Most_Similar Neighbor, 109 of the 126 cohorts are correctly classified (38 of the 46 controls, 47 of the 50 IC patients, and 24 of the 30 BC patients). Again, the quality of this procedure decreases as the number of neighbors increases.

These simple examinations show that there are features in these spectra that separate these cohorts to some degree since the results for four neighbors still yield classifications that are significantly above those expected from random chance (36.51, 39.68, and 23.81% for control, IC patients, and BC patients, respectively). By using a Feature Selection method to search for the optimum set of J features, a model using four nearest neighbors should yield results that are superior to the 4-Neighbor results listed in Tables 1(A) and (B). Including the distance dependence will cause the classification of a point to reflect the local environment in this J-dimensional space (i.e. proximity of neighbors and their classifications), and may increase or decrease the accuracy of the classifier.

The first set of classification results uses four features and a Euclidean distance. In all runs, the EP Feature Selection method has a population size of 2000 (sets of four features) and the search runs for 400 generations. The intensity change between two cohorts in a given bin can be either a relative difference (i.e., the absolute difference divided by their average distance)

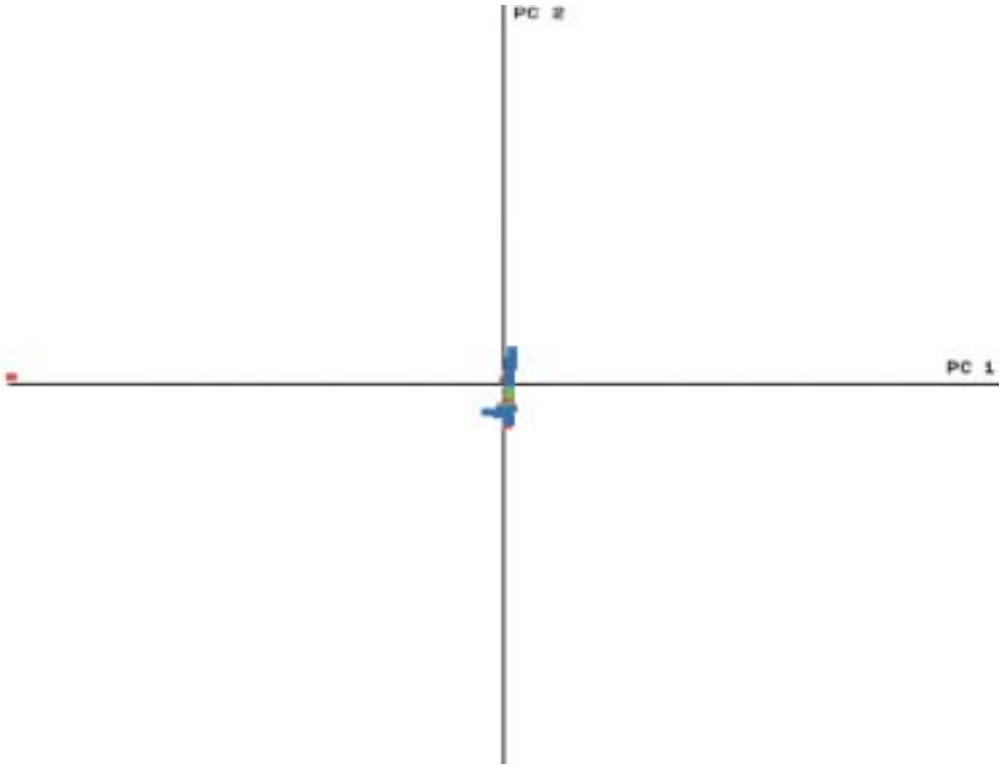

Fig. 3. Principle Component analysis of the 127 ^{1}H-NMR spectra obtained of normal controls, IC patients and BC patients showing the presence of an outlier. This outlier was also confirmed using a two-dimensional Sammon Map (data not shown).

or an absolute difference. An inverse probability function is used throughout and the probability of being in the Unknown-Class has a maximum value of 0.1 for all neighbors. The value of HALF can be 0.1, 0.15, and 0.2, representing increasing widths in the probability function. The results for the six runs are shown in Table 2. Included in this table are the results for the best set of features and for the 50th best set. In addition, the features used in each of the best 50 sets are examined, and if a feature is used in five or more sets it is listed along with the number of times it appears in these sets.

These results show that selecting an optimal set of four features produces better classification models than the one using the overall NMR profile ($K = 4$ result in Table 1(A)). It is interesting to note that greater accuracy is obtained when the absolute difference between intensities is used and that this accuracy is less affected by changing the value of HALF (α). Conversely, there is a larger drop in the overall accuracy when the 50th best feature set is compared to the best, but this decrease is less than 4%. The features present in the top 50 sets do not significantly change when HALF changes, but become very different when the intensity change is either a relative or absolute difference.

Very similar results are obtained when a Manhattan distance is used instead of a Euclidean distance (Table 3), though the overall accuracy of the best feature set increases $\sim$1% when the relative difference is used and $>$1% when the absolute difference is used. The most heavily used features in the top 50 sets in Table 2 are still the most heavily used when a Manhattan distance is applied (Table 3), though there are some changes in the less-used features.

The predicted classifications for the 126 cohorts using the best feature set from run KNN(4b5) are shown in Table 4. The source (i.e. Class) of the urine sample (i.e. normal, IC patient, or BC patient) is also shown. These results show that in the great majority of cases the classification is correct and definitive, or there is an obvious question about the classification. For example, the first two cohorts are almost evenly assigned to normal healthy and IC patient, so their classification is undeterminable between these two classes. The next two cohorts have a high probability of being from an IC patient, but have a 14.4 and 27.9% chance of being unknown. This result means that they are quite far from two or more of their four neighbors and the confidence in the classification is reduced. In only a few cases

Table 2
Probability of correct classification when four features within the ^{1}H-NMR spectra are used with a Euclidean distance in a Distance-Dependent Four-Nearest Neighbor study

Parameter	knn4a1	knn4a2	knn4a3	knn4a4	knn4a5	knn4a6
Intensity change	Diff/Avg	Diff/Avg	Diff/Avg	Diff	Diff	Diff
HALF	0.1	0.15	0.2	0.1	0.15	0.2
Best accuracy	78.35%	80.35%	81.21%	83.03%	83.29%	83.46%
Normal controls	80.48%	83.43%	82.36%	87.91%	88.44%	88.79%
IC patients	86.34%	87.29%	88.51%	87.00%	87.00%	87.00%
BC patients	61.77%	64.06%	67.29%	68.91%	69.22%	69.39%
50th accuracy	76.86%	79.26%	79.95%	79.48%	79.84%	79.99%
Normal controls	76.28%	79.02%	79.98%	84.56%	84.98%	82.26%
IC patients	86.18%	74.92%	85.29%	84.04%	83.55%	88.06%
BC patients	62.24%	70.21%	71.00%	64.09%	65.77%	63.05%
Feature 189	42	40	40	6	5	5
Feature 332	0	0	0	14	14	13
Feature 352	1	1	1	8	8	8
Feature 377	43	41	43	0	0	0
Feature 383	7	10	8	36	35	34
Feature 384	0	0	0	4	6	6
Feature 437	5	4	4	39	39	40
Feature 470	43	40	41	0	0	0
Feature 483	5	3	3	46	44	45

Table 3
Probability of correct classification when four features within the ^{1}H-NMR spectra are used with a Manhattan distance in a Distance-Dependent Four-Nearest Neighbor study

Parameter	knn4b1	knn4b2	knn4b3	knn4b4	knn4b5	knn4b6
Intensity change	Diff/Avg	Diff/Avg	Diff/Avg	Diff	Diff	Diff
HALF	0.1	0.15	0.2	0.1	0.15	0.2
Best accuracy	79.58%	81.44%	82.31%	83.73%	83.96%	84.10%
Normal controls	81.54%	84.43%	85.83%	89.07%	89.57%	89.86%
IC patients	87.66%	88.52%	89.00%	88.52%	88.52%	88.52%
BC patients	63.14%	65.06%	65.78%	67.57%	67.76%	67.90%
50th accuracy	78.22%	79.67%	80.22%	79.79%	80.03%	80.14%
Normal controls	77.34%	79.37%	79.46%	82.47%	78.38%	84.02%
IC patients	85.83%	86.64%	87.12%	89.92%	89.19%	86.05%
BC patients	66.89%	68.52%	69.87%	58.81%	67.28%	64.34%
Feature 189	45	39	39	2	2	2
Feature 356	1	5	4	1	1	1
Feature 377	41	41	42	1	1	1
Feature 383	5	9	7	44	44	46
Feature 437	5	7	7	47	46	47
Feature 470	45	40	39	0	0	0
Feature 483	3	3	3	49	48	49

(Cohorts 101 and 122, for example) does this model give a definitively wrong classification. Considering 0.500 as a threshold for the correct classification of any particular sample, the results show that 93.5% (43 our of 46) of the normal controls are correctly classified, while 92% (46 out of 50) and 73.3% (22 out of 30) of the IC and BC patient samples are correctly classified. Twenty percent (6 out of 30) of the BC samples were misclassified as IC samples, while only 6.7% were misclassified as normal controls (2 out of 30). For the misclassified IC samples, there was little difference in their rate of misclassification as either normal controls (4%), BC patients (2%), or indeterminable (2%). Again, if 0.500 is considered as a threshold for a correct classification, the sensitivity and specificity for the diagnosis of the IC patients is 92% (46 out of 50 correctly classified as IC) and 89.5% (8 out of 76 misclassified as IC). The sensitivity and specificity for the diagnosis of BC is 73.3% (22 out of 30 correctly classified as IC) and 99.0% (1 out of 96 misclassified as IC) and for normal controls the sensitivity and specificity is 93.5% (43 out of 46 correctly classified as a normal control) and 95%

Table 4
Experimental and probabilistic classification of the 126 cohorts produced by the best set of four features from run KNN(4b5) shown in Table 3 (i.e. 370, 383, 437, and 483)

Sample number	Class	Probabilistic classification			
		Normal	IC	BC	Unknown
1	Normal	0.551	0.442	0.000	0.007
2	Normal	0.435	0.550	0.000	0.015
3	Normal	0.856	0.000	0.000	0.144
4	Normal	0.721	0.000	0.000	0.279
5	Normal	1.000	0.000	0.000	0.000
6	Normal	1.000	0.000	0.000	0.000
7	Normal	1.000	0.000	0.000	0.000
8	Normal	1.000	0.000	0.000	0.000
9	Normal	1.000	0.000	0.000	0.000
10	Normal	1.000	0.000	0.000	0.000
11	Normal	1.000	0.000	0.000	0.000
12	Normal	1.000	0.000	0.000	0.000
13	Normal	1.000	0.000	0.000	0.000
14	Normal	0.557	0.158	0.286	0.000
15	Normal	0.685	0.164	0.151	0.000
16	Normal	1.000	0.000	0.000	0.000
17	Normal	1.000	0.000	0.000	0.000
18	Normal	1.000	0.000	0.000	0.000
19	Normal	1.000	0.000	0.000	0.000
20	Normal	1.000	0.000	0.000	0.000
21	Normal	1.000	0.000	0.000	0.000
22	Normal	0.545	0.189	0.000	0.266
23	Normal	1.000	0.000	0.000	0.000
24	Normal	1.000	0.000	0.000	0.000
25	Normal	1.000	0.000	0.000	0.000
26	Normal	1.000	0.000	0.000	0.000
27	Normal	1.000	0.000	0.000	0.000
28	Normal	1.000	0.000	0.000	0.000
29	Normal	1.000	0.000	0.000	0.000
30	Normal	0.445	0.332	0.224	0.000
31	Normal	1.000	0.000	0.000	0.000
32	Normal	0.777	0.000	0.223	0.000
33	Normal	1.000	0.000	0.000	0.000
34	Normal	1.000	0.000	0.000	0.000
35	Normal	1.000	0.000	0.000	0.000
36	Normal	1.000	0.000	0.000	0.000
37	Normal	1.000	0.000	0.000	0.000
38	Normal	1.000	0.000	0.000	0.000
39	Normal	1.000	0.000	0.000	0.000
40	Normal	1.000	0.000	0.000	0.000
41	Normal	1.000	0.000	0.000	0.000
42	Normal	0.823	0.177	0.000	0.000
43	Normal	1.000	0.000	0.000	0.000
44	Normal	0.000	1.000	0.000	0.000
45	Normal	1.000	0.000	0.000	0.000
46	Normal	0.807	0.193	0.000	0.000
47	IC	0.000	1.000	0.000	0.000
48	IC	0.000	1.000	0.000	0.000
49	IC	0.000	1.000	0.000	0.000
50	IC	0.000	1.000	0.000	0.000
51	IC	0.000	1.000	0.000	0.000
52	IC	0.000	1.000	0.000	0.000
53	IC	0.000	0.685	0.315	0.000
54	IC	0.000	1.000	0.000	0.000
55	IC	0.000	1.000	0.000	0.000

Table 4, continued

Sample number	Class	Probabilistic classification			
		Normal	IC	BC	Unknown
56	IC	0.000	1.000	0.000	0.000
57	IC	0.000	1.000	0.000	0.000
58	IC	0.000	1.000	0.000	0.000
59	IC	0.000	1.000	0.000	0.000
60	IC	0.000	1.000	0.000	0.000
61	IC	0.000	1.000	0.000	0.000
62	IC	0.000	1.000	0.000	0.000
63	IC	0.093	0.907	0.000	0.000
64	IC	0.995	0.000	0.000	0.005
65	IC	0.000	0.915	0.085	0.000
66	IC	0.000	1.000	0.000	0.000
67	IC	0.000	1.000	0.000	0.000
68	IC	0.000	1.000	0.000	0.000
69	IC	0.000	1.000	0.000	0.000
70	IC	0.000	0.914	0.086	0.000
71	IC	0.000	1.000	0.000	0.000
72	IC	0.000	0.213	0.787	0.000
73	IC	0.000	1.000	0.000	0.000
74	IC	0.000	1.000	0.000	0.000
75	IC	0.000	0.700	0.300	0.000
76	IC	0.000	1.000	0.000	0.000
77	IC	0.057	0.943	0.000	0.000
78	IC	0.052	0.948	0.000	0.000
79	IC	0.056	0.944	0.000	0.000
80	IC	0.000	1.000	0.000	0.000
81	IC	0.000	1.000	0.000	0.000
82	IC	0.094	0.811	0.095	0.000
83	IC	0.000	1.000	0.000	0.000
84	IC	0.516	0.247	0.237	0.000
85	IC	0.000	1.000	0.000	0.000
86	IC	0.324	0.448	0.228	0.000
87	IC	0.000	1.000	0.000	0.000
88	IC	0.000	1.000	0.000	0.000
89	IC	0.000	0.842	0.158	0.000
90	IC	0.206	0.558	0.236	0.000
91	IC	0.000	0.846	0.154	0.000
92	IC	0.237	0.763	0.000	0.000
93	IC	0.000	1.000	0.000	0.000
94	IC	0.000	1.000	0.000	0.000
95	IC	0.000	0.788	0.212	0.000
96	IC	0.214	0.786	0.000	0.000
97	BC	0.000	0.000	1.000	0.000
98	BC	0.330	0.000	0.670	0.000
99	BC	0.000	0.164	0.836	0.000
100	BC	0.500	0.224	0.276	0.000
101	BC	0.000	1.000	0.000	0.000
102	BC	0.000	0.151	0.849	0.000
103	BC	0.000	0.162	0.838	0.000
104	BC	0.000	0.080	0.845	0.074
105	BC	0.000	0.085	0.844	0.071
106	BC	0.000	0.882	0.000	0.118
107	BC	0.000	0.000	1.000	0.000
108	BC	0.000	0.000	1.000	0.000
109	BC	0.000	0.000	1.000	0.000
110	BC	0.719	0.000	0.281	0.000
111	BC	02.32	0.786	0.000	0.000
112	BC	0.000	0.000	1.000	0.000
113	BC	0.000	0.000	1.000	0.000

Table 4, continued

Sample number	Class	Probabilistic classification			
		Normal	IC	BC	Unknown
114	BC	0.000	0.000	1.000	0.000
115	BC	0.000	0.000	1.000	0.000
116	BC	0.000	0.214	0.786	0.000
117	BC	0.000	0.000	1.000	0.000
118	BC	0.000	0.000	1.000	0.000
119	BC	0.000	0.000	1.000	0.000
120	BC	0.000	0.430	0.570	0.000
121	BC	0.000	0.495	0.505	0.000
122	BC	0.000	1.000	0.000	0.000
123	BC	0.218	0.000	0.782	0.000
124	BC	0.000	0.000	1.000	0.000
125	BC	0.000	0.754	0.246	0.000
126	BC	0.231	0.769	0.000	0.000

(4 out of 80 misclassified as a normal control).

A graphical analysis of the features presented in Tables 2 and 3 produce a few interesting results. A display of the intensity of feature 377, which is used prominently in the top feature sets when a relative difference in intensities is used, is shown in Fig. 4(A). In this plot, the 46 normal control individuals are shown in the first set of green data points, the 50 IC patients are the red points, and the 30 BC patients are the blue. The dotted lines represent the average intensities for each of the three Classes. Though the average is exaggerated by two high intensity values, the average intensity of this feature is greater in normal control individuals than in either the IC or BC patients.

Similar plots are shown for Features 356, 384, and 437 in Figs 4(B–D), respectively. They also show a reasonable separation between the normal control individuals and the IC- or BC-affected patients, and it is interesting to note that Feature 437 is present in five or more of the best sets for at least one run using either intensity change measure or either distance metric. Though feature 356 is only present five times in one of the 12 runs listed in Tables 2 and 3, it is the only feature of those listed that shows a reasonable separation in the averages of all three cohorts.

When eight features are used instead of four, the results obtained using Euclidean and Manhattan distances are shown in Tables 5 and 6, respectively. Because the search space of eight unique feature sets is many orders of magnitude larger than for four feature sets, the EP method uses a population size of 4000 and runs for 800 generations.

These results again show that an absolute difference in intensities produces better classifiers than relative differences and that the former is less sensitive to changes in HALF. What these results also show is that as the number of features increases from four to eight the overall quality if the classifier is virtually independent upon whether Euclidean or Manhattan distances are used. Euclidean distances generally improve the classification of BC-affected patients, while the Manhattan distance classifiers generally improve the classification of control individuals and IC-affected patient cohorts. The use of more features should continue to improve the accuracy of the classification model up to a certain point. A 100% correct classification would not be expected in a four-neighbor model for all cohorts, and increasing the accuracy more would require polling less than four neighbors.

4. Discussion

The development of technologies that provide a global view of the cell at the genomic, transcriptomic, proteomic, and metabonomic level is and will continue to be a major trend in biological science for the foreseeable future. While there are many different types of information that can be gleaned using these global approaches, one of the major initiatives is to use these technologies to more effectively diagnose diseases and develop better therapies. A vast majority of these initiatives use these technologies to rapidly screen thousands of species within complex mixtures in search of a biomarker that is unique to either the healthy or diseased state. The present approach, however, does not rely on a single unique species, but rather it takes into account the abundances of several key features within each spectrum to select the diagnosis.

An effective diagnostic tool should be, amongst other things, non-invasive, technically simple, and require a minimal amount of sample. While such tools currently exist to screen urine samples for evidence of acute BC (including urine dipstick and microscopy) the current "diagnosis" of IC often involves cystoscopy with hydrodistension performed under general anesthesia. Unfortunately, findings of glomerulations or Hunner's ulcers at cystoscopy with hydrodistension does not even provide a definitive diagnosis of IC, however, it is recommended for fulfilling NIDDK diagnostic criteria for IC. Although other diagnostic parameters (including the measurement of urine APF activity or HB-EGF/EGF levels) have been described for IC, a key advantage of using NMR-based technology is that many of the resonances observed in a typical spectrum of any human biofluid may be readily assignable to a known compound based solely on the resonance frequency values, thereby potentially providing additional informa-

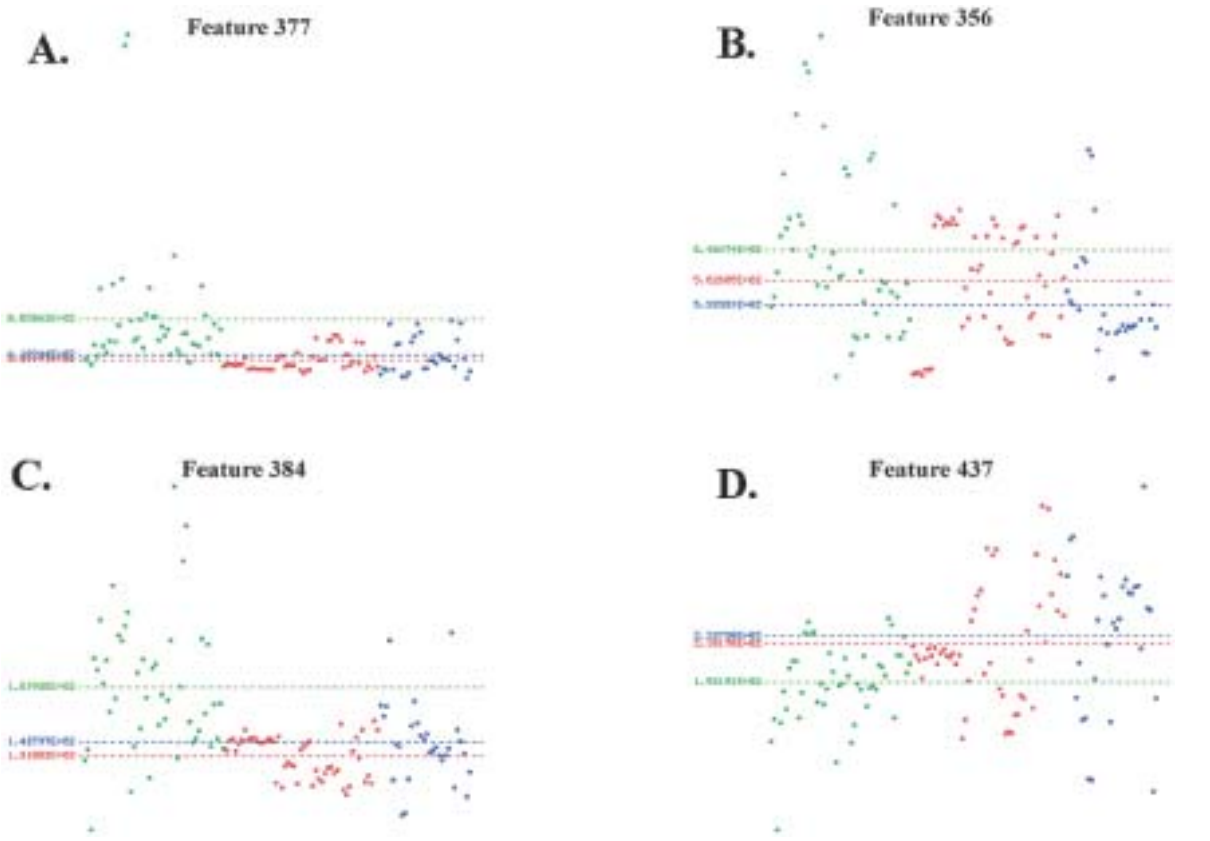

Fig. 4. A display of the intensity of features (A) 377, (B) 356, (C) 384, and (D) 437 which are used prominently in the top feature sets used in the diagnosis of the various conditions (i.e. normal controls vs. IC vs. BC).

tion about the disease process itself. In addition, for those signals that cannot be readily assigned, experiments such as total correlated spectroscopy (TOCSY), correlated spectroscopy (COSY), nuclear Overhauser spectroscopy (NOESY), etc. can be used in an attempt to identify their compounds of origin. Clearly, the relative sensitivity, specificity, positive predictive value, negative predictive value, and cost for each type of analysis will need to be considered for determining the optimal diagnostic test for IC.

An obvious concern in using pattern matching of spectra generated from biofluids for disease diagnostics is the variability of the samples from the human subjects. Unlike experimental animals, such as mice, humans cannot be kept under strictly controlled conditions of diet, rest, physical activity, or drug intake (especially for over the counter medications). While there is no universally accepted treatment for IC, many of the individuals affected by IC in this study were using a variety of different medications with the goal of alleviating their symptoms. While some patients were on no medications, most were on various medications including pentosan polysulfate (Elmiron), dimethyl sulfoxide (DMSO), aloe vera, nonsteroidal anti-inflammatory medications, and antihistamines. To show that the ^{1}H-NMR-based diagnostic was not simply classifying spectra based on the medications each individual was taking, patients that formed each node in the diagnostic pattern were analyzed based on their medication intake. It was found that the cluster the individuals fell into was independent of their drug intake. For example, IC-affected individuals taking no medication were spread out amongst the various clusters that were diagnostic of IC and individuals taking medications were also distributed within the various IC-clusters as well. In addition, none of the frequency values so far identified as being key to generating the diagnostic patterns were directly related to the medication being taken by the individual or any of their known metabolites.

The MS-based analysis was able to correctly classify the urine samples as being obtained from either normal or IC-affected individuals with an accuracy of 100%. However, the bioinformatic tool used to segregate the spectra is unable to perform a three-tiered

Table 5
Probability of correct classification when eight features within the ^{1}H-NMR spectra are used with a Euclidean distance in a Distance-Dependent Four-Nearest Neighbor study

Parameter	knn8a1	knn8a2	knn8a3	knn8a4	knn8a5	knn8a6
Metric	Diff/Avg	Diff/Avg	Diff/Avg	Diff	Diff	Diff
HALF	0.1	0.15	0.2	0.1	0.15	0.2
Best accuracy	80.24%	84.97%	87.28%	87.64%	88.02%	88.23%
Normal controls	75.18%	83.46%	87.58%	92.73%	93.27%	92.70%
IC patients	89.06%	88.59%	89.33%	88.88%	89.10%	90.45%
BC patients	73.29%	81.26%	83.41%	77.77%	78.15%	77.54%
50th accuracy	79.93%	84.40%	86.53%	87.16%	87.53%	87.75%
Normal controls	78.44%	82.81%	86.09%	92.81%	92.97%	92.37%
IC patients	87.49%	87.54%	88.63%	90.28%	89.05%	90.85%
BC patients	69.61%	81.59%	83.70%	73.28%	76.63%	75.50%

Table 6
Probability of correct classification when eight features within the ^{1}H-NMR spectra are used with a Manhattan distance in a Distance-Dependent Four-Nearest Neighbor study

Parameter	knn8b1	knn8b2	knn8b3	knn8b4	knn8b5	knn8b6
Metric	Diff/Avg	Diff/Avg	Diff/Avg	Diff	Diff	Diff
HALF	0.1	0.15	0.2	0.1	0.15	0.2
Best accuracy	82.56%	85.36%	86.59%	87.78%	88.10%	88.25%
Normal controls	85.23%	88.33%	88.48%	91.58%	92.02%	92.17%
IC patients	87.38%	89.54%	90.67%	92.18%	92.35%	92.52%
BC patients	70.45%	73.84%	76.90%	74.61%	74.99%	75.12%
50th accuracy	82.17%	84.87%	86.04%	87.00%	87.36%	87.53%
Normal controls	84.77%	87.61%	89.54%	90.53%	90.83%	92.62%
IC patients	88.00%	90.28%	90.52%	90.49%	92.52%	90.39%
BC patients	68.45%	71.65%	73.19%	75.78%	73.42%	74.96%

classification. Therefore, an alternative bioinformatic analysis was used to determine if urine samples from three different conditions could be correctly diagnosed. An analysis of the 531-binned NMR spectra of 126 cohorts produced a single model that is able to correctly classify the cohorts to approximately 84% level. It uses a Distance-Dependent Four-Nearest Neighbor procedure to predict the Class of each cohort, and the resulting distribution of Class probabilities can suggest to the researcher that the classification of a particular cohort is suspect. The classification of the control and IC patient cohorts is more accurate than the BC patient cohorts, and this may be caused by either the smaller size of the BC patients training set and/or the lack of a strong biomarker specific to the diagnosis of BC. However, NMR would never be cost-effective enough for the diagnosis of BC as there are fairly cheap, sensitive and specific ways to diagnose BC now that NMR could never compete with.

Acknowledgements

This project has been funded in whole or in part with Federal funds from the National Cancer Institute, National Institutes of Health, under Contract No. NO1-CO-12400 (TDV), grant NIDDK R01 DK52596 (SK) and VA Merit Review funding (SK). We also wish to thank Dr. Bruce Adams from Varian, Inc. for the NMR binning package.

References

[1] J.T. Brindle, H. Antti, E. Holmes, G. Tranter, J.K. Nicholson, H.W. Bethell, S. Clarke, P.M. Schofield, E. McKilligin, D.E. Mosedale and D.J. Grainger, Rapid and noninvasive diagnosis of the presence and severity of coronary heart disease using ^{1}H-NMR-based metabonomics, *Nat Med* **8**(12) (Dec. 2002), 1439–1444.

[2] G.C. Curhan, F.E. Speizer, D.J. Hunter, S.G. Curhan and M.J. Stampfer, Epidemiology of interstitial cystitis: a population based study, *J Urol* **161** (1999), 549–552.

[3] Division of Kidney, Urologic, and Hematologic Diseases (DKUHD) of the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK). Diagnostic criteria for research studies (interstitial cystitis), *Am J Kidney Dis* **13** (1989), 353–354.
[4] D.R. Erickson, S.X. Xie, V.P. Bhavanandan, M.A. Wheeler, R.E. Hurst, L.M. Demers, L. Kushner and S.K. Keay, A comparison of multiple urine markers for interstitial cystitis, *J Urol* **167** (2002), 2461–2469.
[5] J.H. Holland, ed., *Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence,* Third edition, MIT Press: Cambridge, MA, USA, 1994.
[6] S.L. Johansson and M. Fall, Clinical features and spectrum of light microscopic changes in interstitial cystitis, *J Urol* **143** (1990), 1118–1124.
[7] S. Keay, C-O. Zhang, M. Hise, A.L. Trifillis, J.R. Hebel, S.C. Jacobs and J.W. Warren, Decreased 3H-thymidine incorporation by human bladder epithelial cells following exposure to urine from interstitial cystitis patients, *J Urol* **156** (1996), 2073–2078.
[8] S. Keay, C-O. Zhang, D.I. Kagen, M.K. Hise, S.C. Jacobs, J.R. Hebel, D. Gordon, K. Whitmore, S. Bodison and J.W. Warren, Concentrations of specific epithelial growth factors in the urine of interstitial cystitis patients and controls, *J Urol* **158** (1997), 1983–1988.
[9] S. Keay, C-O. Zhang, M.K. Hise, J.R. Hebel, S.C. Jacobs, D. Gordon, K. Whitmore, S. Bodison, N. Gordon and J.W. Warren, A diagnostic *in vitro* assay for interstitial cystitis, *Urology* **52**(6) (1998), 974–978.
[10] S. Keay, C-O. Zhang, J. Shoenfelt, D.R. Erickson, K. Whitmore, J.W. Warren, R. Marvel and T. Chai, Sensitivity and specificity of antiproliferative factor, heparin-binding epidermal growth factor-like growth factor, and epidermal growth factor as urine markers for interstitial cystitis, *Urology* **57**(6 Suppl 1) (2001), 9–14.
[11] T. Kohonen, The self-organizing map, *Proc. Inst. Electrical Electronics Eng.* **78** (1990), 1464–1480.
[12] J.C. Lindon, J.K. Nicholson, E. Holmes and J.R. Everett, *Prog. NMR Spectrosc.* **12** (2000), 289–320.
[13] J.K. Nicholson, J.C. Lindon and E. Holmes, *Xenobiotica* **29** (1999), 1181–1189.
[14] J.K. Nicholson, J. Connelly, J.C. Lindon and E. Holmes, *Nat. Drug. Discov.* **1** (2002), 153–161.
[15] K.J. Oravisto, Epidemiology of Interstitial Cystitis, *Ann Chir Gyn Fenn* **64** (1975), 75–77.
[16] E.F. Petricoin, III et al., Use of proteomic patterns in serum to identify ovarian cancer, *The Lancet* **359** (2002), 572–577.
[17] T.L. Ratliff, C.G. Klutke and E.M. McDougall, The Etiology of Interstitial Cystitis, *Urol Clin N Am* **21** (1994), 21–29.
[18] M.R. Ruggieri, M.J. Chelsky, S.I. Rosen, T.J. Shickley and P.M. Hanno, Current findings and future research avenues in the study of interstitial cystitis, *Urol Clin North Am* **21** (1994), 163–176.
[19] S.H. Smallcombe, S.L. Patt and P.A. Keifer, *J. Magn. Reson. A* **117** (1995), 295–303.
[20] B.H. Smith and L.P. Dehner, Chronic ulcerating interstitial cystitis (Hunner's ulcer), *Arch Path* **93** (1972), 76–81.
[21] J.A. Waxman, P.J. Sulak and T.J. Kuehl, Cystoscopic findings consistent with interstitial cystitis in normal women undergoing tubal ligation, *J Urol* **160** (1998), 1663–1667.