Adaptive Learning of Immunosignaturing Features for Multi-Disease Pathologies

Anna Malin\textsuperscript{1}, Narayan Kovvali\textsuperscript{1}, Antonia Papandreou-Suppappola\textsuperscript{1}, Brian O’Donnell\textsuperscript{1}, Stephen Johnston\textsuperscript{2}, Phillip Stafford\textsuperscript{2}

\textsuperscript{1}School of Electrical, Computer and Energy Engineering, Arizona State University, Tempe, AZ
\textsuperscript{2}The Biodesign Institute, Arizona State University, Tempe, AZ

 Abstract—Previously, adaptive learning algorithms have been used with immunosignaturing in order to identify disease states in patients. However, in these algorithms the presence of a single disease state is assumed, although in a clinical setting this may not be the case. We propose a novel algorithm based on latent feature identification using beta process factor analysis, in which the binary feature sharing matrix is modified and key comparisons are applied to identify multiple possible underlying disease states. The algorithm is verified using combinations of actual patient immunosignaturing data. The proposed method has a variety of applications, including multi-disease state diagnosis in the clinical setting, multi-biothreat detection in the field, and separation of co-contaminated biological samples.

I. INTRODUCTION

In a clinical setting, the diagnosis of patients with multiple disease states can be difficult in comparison with single disease diagnosis. Additional tests, procedures, and consultations may be required in order to distinguish the various disease states based on a variety of sometimes conflicting symptoms. As such, it is desired to have a broad based test that is capable of identifying multiple disease states in patients, without the need for repeated, serial testing. Immunosignaturing is one technique that has been previously identified as a capable method for single disease clustering [1, 2]. Immunosignaturing uses random-sequence peptide microarrays to analyze antibodies in patient samples [3]–[5]. Peptide sequences are plated onto a microarray slide, and the antibodies present in the sample preferentially bind with peptide sequences that have favorable organic chemistry interactions [6]. The strength of antibody binding is quantified by fluorescence measurements and the median intensity measurement for each spot is recorded. Thousands of random peptide sequences are used on the microarray to analyze a small quantity of patient sample, such as blood or saliva.

A variety of methods have been developed for microarray data analysis, including ANOVA [7], PCA [4], [8], [9], and clustering [10]–[13]. In [5], several clustering methods were compared for immunosignaturing, with a naive Bayes approach showing the best results. These methods are not adaptive and require reanalysis every time additional patient data is received. Adaptive methods were investigated in [1] and [2], which allow for active model adaptation and classification based on continuously obtained patient data. While these methods, along with [5], have shown promising results for immunosignaturing analysis and clustering, the underlying assumption was that a patient only suffers from a single disease pathology at any given time. These methods cannot be used without modification when generalization to multiple pathologies is required. At present there are no works that explicitly address the case where patients may have multiple simultaneous pathologies.

In this paper, we propose a novel immunosignaturing microarray data analysis approach that can be used for identifying multiple disease pathologies in patients. The method is based on beta process factor analysis (BP-FA) [14], [15], which uses Bayesian nonparametric beta process priors to discover latent phenomena, the latent factor model for binding interactions can be modeled using combinations of independent features. In light of known disease states, it is still assumed that the data disease pathology at any given time. These methods cannot be used without modification when generalization to multiple pathologies is required. At present there are no works that explicitly address the case where patients may have multiple simultaneous pathologies.

In this paper, we propose a novel immunosignaturing microarray data analysis approach that can be used for identifying multiple disease pathologies in patients. The method is based on beta process factor analysis (BP-FA) [14], [15], which uses Bayesian nonparametric beta process priors to discover latent phenomena, the latent factor model for binding interactions can be modeled using combinations of independent features. In light of known disease states, it is still assumed that the data disease pathology at any given time. These methods cannot be used without modification when generalization to multiple pathologies is required. At present there are no works that explicitly address the case where patients may have multiple simultaneous pathologies.

II. FEATURE MODEL FOR IMMUNOSIGNATURES

A. BP-FA of Immunosignature Microarray Data

The BP-FA is utilized to learn latent feature combinations underlying the immunosignaturing microarray data. It is important to note that even though multiple disease states are allowed to be present in a patient, and the features are analyzed in light of known disease states, it is still assumed that the data can be modeled using combinations of independent features. While disease pathology and antibody response are complex phenomena, the latent factor model for binding interactions between antibodies corresponding to each disease and the microarray peptide sequences can be used as long as effects...
such as competitive inhibition are not dominant [18], [19]. The BP-FA [14], [15], [20], [21] is a Bayesian nonparametric approach which uses the beta process prior [22], [23] to discover an unlimited number of latent features in data. In immunosignaturing microarray data, these factors can represent underlying patient disease relationships.

The beta process [14], [22] can be defined as the limit $K \to \infty$ of the process:

$$H_K = \sum_{k=1}^{K} q_k \delta_{\phi_k},$$

(1)

where $q_k \sim \text{Beta}(\alpha/K, \beta(K - 1)/K)$ and $\phi_k \sim H_0$, with $\alpha$ and $\beta$ positive scalars and $H_0$ a base distribution. The beta process can be used as a nonparametric prior in the latent factor model

$$X = \Phi Z + E,$$

(2)

where $X$ is the $D \times N$ data matrix comprised of $N$ $D$-dimensional data vectors, $\Phi$ is a $D \times K$ matrix comprised of $K$ $D$-dimensional latent features, $Z$ is a $K \times N$ binary feature sharing matrix, and $E$ is the $D \times N$ measurement error matrix. The BP-FA data model is given by [14]:

$$\phi_k \sim \mathcal{N}(0, \Sigma),$$

$$q_k \sim \text{Beta}\left(\frac{\alpha}{K}, \frac{\beta(K - 1)}{K}\right),$$

$$z_{k,i} \sim \text{Bernoulli}(q_k),$$

$$e_i \sim \mathcal{N}(0, \sigma^2_e I),$$

$$\sigma^2_e \sim \mathcal{I}(c, d),$$

(3)

for $i = 1, \ldots, N$, and $k = 1, \ldots, K$, where $\mathcal{N}$ and $\mathcal{I}$ denote the Gaussian and inverse Gamma distributions, respectively. Note that for simplicity here we use an unweighted version of the model in [14]. The parameters $\alpha$ and $\beta$ may be thought of as tuning parameters that impact how the latent features are shared in the dataset.

For the immunosignaturing microarray data, the data matrix $X$ consists of $D$-dimensional median binding intensity fluorescence measurements (or some dimensionality reduced form thereof) obtained from $N$ patients. The latent feature matrix $\Phi$ models the relationship between the disease antibodies and microarray peptide sequences. While not explored further in this work, this matrix may be used to highlight those antibodies that play a role in various feature combinations. This can be used for further disease analysis, especially if there is a desire to refine further immunosignaturing tests (for example, in reducing the number of microarray points for subsequent studies). The binary matrix $Z$ depicts the relationship between patients and latent features, and is further analyzed in this paper for the identification of possible diseases in the multi-disease patients. Note that in the truncated model a maximum of $K$ latent features can be discovered.

The BP-FA model parameters are estimated using Markov chain Monte Carlo (MCMC) techniques, executed via a Gibbs sampler [16], [17]. The algorithm iteratively updates blocks of parameters by sampling from the following conditional distributions:

$$p(\phi_k | X, \Phi_{-k}, Z, \sigma^2_e) \propto p(X | \phi_k, \Phi_{-k}, Z, \sigma^2_e) p(\phi_k | \Sigma),$$

$$p(q_k | Z) \propto p(Z | q_k) p(q_k | K, \alpha, \beta),$$

$$p(z_{k,i} | x_i, \Phi, z_{-k,i}, \sigma^2_e) \propto p(x_i | z_{k,i}, \Phi, z_{-k,i}, \sigma^2_e) p(z_{k,i} | q_k),$$

$$p(\sigma^2_e | X, \Phi, Z) \propto p(X | \sigma^2_e, \Phi, Z) p(\sigma^2_e | c, d),$$

(4)

for $i = 1, \ldots, N$, and $k = 1, \ldots, K$, where $\Phi_{-k}$ and $z_{-k,i}$ are the corresponding matrix/vector with the $k$th column/element removed. Upon convergence, estimates of the latent feature and feature sharing matrices are computed by averaging the samples.

**B. Disease Identification using Latent Features**

The adaptive learning methods presented in [1], [2] were designed for diagnosis of patients with single disease states and cannot be applied in multi-disease scenarios. We now describe a generalized approach, namely $Z$-matrix PRocessing Enhancement for Immunosignaturing Classification Testing for n states (ZPREDICTn), for the identification of multiple disease states in patients. In this method, the median binding intensity measurements obtained from the immunosignaturing microarray are first processed using principal component analysis (PCA) for dimensionality reduction, with a modified covariance matrix utilized due to high data dimensionality [24]. The resulting reduced-dimension data is then analyzed using the BP-FA to determine the binary latent feature sharing matrix $Z$.

The binary entries of $Z$ indicate the presence or absence of a feature contribution, are characteristic of each disease, and can be compared for matches leading to disease identification. To sufficiently penalize incorrect matches, $Z$ is modified and scaled to obtain a matrix $Y$. In order to determine the underlying disease relationships embedded in the feature sharing matrix in relation to individual disease states, a master key for single disease states is used for comparison against the individual patients with multiple disease states. Note that the binding intensity data corresponding to the keys also undergoes the PCA and BP-FA steps. The matrix $Y$ can be divided into two portions: the test feature vectors representing unknown multiple diseases (matrix $P$) and the key feature vectors representing known single diseases (matrix $N$). The feature vectors in $N$ are weighted and averaged over patients to generate a matrix $M$ consisting of a single representative feature vector for each known disease state (master keys), and $M$ is appended to $P$ to obtain the matrix $A$. The columns of $A$ contain the modified feature vectors of the unknown and known disease states, and are compared by taking inner products to finally obtain the comparison matrix $C$. These
operations can be written as:

\[
\begin{align*}
Y &= g \cdot (2Z - 1), \\
Y &= [N \ P], \\
M_r &= \sum p_{q(r)} N_{q(r)}, \quad r = 1, \ldots, R, \\
A &= [M \ P], \\
C &= UT[(hA)^T(hA), R],
\end{align*}
\]

where \(g\) and \(h\) are scaling factors, \(p_{q(r)}\) are weighting coefficients for the known disease state feature vectors, \(R\) is the number of known disease states, and \(UT[\cdot, R]\) signifies keeping only the upper triangular portion of the matrix that correspond to the desired \(R\) columns. In general, the weighting coefficients \(p_{q(r)}\) can be adjusted to account for varying amounts of competing antibodies and to incorporate information surrounding inhibition, such as competitive inhibition [18], [19]. The block diagram of the ZPREDICTn algorithm is shown in Figure 1.

III. SIMULATION RESULTS

In this section, we present the multi-disease identification results from using the ZPREDICTn algorithm on real immunosignaturing data. The BP-FA implementation was also verified using synthetic data [2]. In order to test the ZPREDICTn algorithm and evaluate its performance, multi-disease datasets are created from real immunosignaturing datasets obtained from experiments with a 10,400 spot random-sequence peptide microarray. Multi-disease immunosignatures are created by averaging together immunosignatures from at least two distinct diseases. This is expected to represent a simplified scenario of multi-disease pathologies where more complex non-additive antibody bindings are not considered. It can also represent practical scenarios where multiple laboratory samples have been accidentally co-contaminated. The master keys for each disease state were created from averages of at least five patient immunosignatures. Where possible (with the availability of data), master key data was created using immunosignatures different from those used to create the multi-disease patient population.

Two different datasets are considered. The first dataset consists of 10 combined multi-disease patient immunosignatures, with the first 5 containing averages of Breast Cancer and Sarcoma and the last 5 containing averages of Sarcoma and Glioma. For the analysis of this dataset 20 key immunosignatures are available, 5 from each of the 4 disease states: Breast Cancer, Sarcoma, Glioma, and Normal. This dataset construction is used to test the ability of the method to identify multiple diseases when the key diseases may or may not correspond to the present diseases. The second dataset consists of 20 combined multi-disease patient immunosignatures with averages from Alzheimer’s and Myeloma. For the analysis of this dataset 59 key immunosignatures are available, with 20 from Alzheimer’s, 20 from Cocci, 10 from Myeloma, and 9 from Normal. This dataset construction is used to determine if the same behavior is observed in larger datasets with more patient diversity.

Following the ZPREDICTn algorithm steps, PCA is first applied to the raw immunosignature median binding intensity values for dimensionality reduction. The number of principal components retained is 11 (corresponding to 93.2% variance) for dataset 1 and 15 (corresponding to 94.7% variance) for dataset 2. The reduced-dimension PCA output for both the patients and keys is next fed into the BP-FA procedure for latent feature discovery. The model truncation limit is set to \(K = 50\), and 2000 burn-in and 2000 sample generation Gibbs sampling iterations are carried out, for both datasets. Finally, the \(Z\) matrix modification calculations are performed and the matrices \(A\) and \(C\) computed. The weighting and scaling parameters are set to \(p_{q(r)} = 0.5\), \(g = 10\), and \(h = 0.1\) for both datasets.

Figure 2 shows plots of the entries of \(A\) (transposed, so that the features are along the x-axis and patients are along the y-axis) and \(C\) (in a stem-plot, with patients along the x-axis and colors representing comparisons with each of the known disease keys) for dataset 1. Additionally, a vector of zeros has been inserted in the plot of \(A\) to clearly depict the two portions: the test feature vectors representing unknown multiple diseases (\(P\)) and the master key feature vectors representing known single diseases (\(M\)). Figure 3 shows corresponding plots of the entries of \(A\) (transposed) and \(C\) (stem-plot) for dataset 2.

From the stem plots of Figures 2(b) and 3(b), the best matching diseases found for each patient using ZPREDICTn can be seen to agree well with the actual multi-disease states. For example, for patient 1 in dataset 2 Alzheimer’s and Myeloma yield the highest inner product values, which is expected as all patients in that dataset have both Alzheimer’s and Myeloma immunosignatures. Thus, while the method does not indicate how many diseases are present in an individual, it can help to identify those that are most likely present.

The disease identification performance of the algorithm is quantified using the true positive and true negative classification rates. The true positive classification rate indicates the cases where patients are correctly identified as having a particular disease and the true negative classification rate indicates the cases where patients were correctly identified as not having a particular disease. Additionally, identification performance is quantified using type 1 and type 2 errors. Type
I errors are false positives (incorrect indication of a disease that a patient does not actually have) and type 2 errors are false negatives (failure to detect a disease present in a patient). For classification, in this work we assume that the number of diseases in each patient ($n = 2$) is known. While it is possible to perform further analysis to determine $n$, that is outside the scope of this work.

The correct classification and error rates showing the performance of the ZPREDICTn algorithm for immunosignaturing dataset 1 and dataset 2 are given in Table I and Table II, respectively. It can be seen from the results that the ZPREDICTn algorithm determines the diseases present in each patient with good but not perfect accuracy. The average correct classification rate is 95% for dataset 1 and 87.5% for dataset 2. Thus, this method is expected to be useful when patients are suspected of having multiple diseases, or in cases when laboratory samples may be co-contaminated.

### IV. CONCLUSION

In this paper, we described a novel immunosignaturing microarray data analysis approach that can be used for identifying
multiple disease pathologies in patients. Feature reduction is first performed on the median binding intensity measurements using PCA, followed by adaptive latent feature determination via the BP-FA, and subsequent analysis is carried out using modified Z matrix comparisons based on a set of known key single disease immunosignatures. Simulation results on averaged combined real patient multi-disease immunosignaturing data demonstrated average true positive and true negative classification rates of at least 87.5% and an average error rate of at most 12.5%, indicating that the proposed ZPREDICTn method may be a good initial step for identifying and classifying multi-disease state samples of immunosignaturing data.

The method is general and can be applied to any number of diseases (parameter \( n \)), assuming that some additional patient information is available. The determination of \( n \) using either prior knowledge or further analysis of the binding intensity data is the subject of future work. Further work is also warranted for extension of the method to more complex biological models, for example by use of a weighted latent factor representation.

REFERENCES