A Data-driven Biomarker Computational Model for Lung Disease Classification

David Gnabasik and Gita Alaghband
Computer Science and Engineering, University of Colorado Denver
Denver, CO USA
(David.Gnabasik, Gita.Alaghband)@ucdenver.edu

Abstract

We develop a data-driven computational model that reliably classifies individual patient into one of 7 non-overlapping lung disease clinical types within our dataset: healthy non-smokers, smokers diagnosed with and without chronic obstructive pulmonary disease (COPD), adenocarcinoma, squamous cell carcinoma, cystic fibrosis, and acute lung injury. Panels of 12 cytokine blood serum biomarker measurements precisely classify both known and unknown patients into one of these distinct clinical types. Our model classifies clinical types and patients directly from the conditional relationships of noisy, incomplete, and variable protein concentration measurements, including outliers. Biomarker concentration measurements induce discrete state variables through a binning algorithm that exposes the conditional relationships and dependencies among the concentration data. A unique application of an XOR operation on the state space extracts the patterns identifying the set of distinctive features for each clinical type. Our model builds a discrete topological structure from a baseline data set, and is developed using several novel schemes designed specifically for this analysis. The result is a multidimensional space representing a characteristic set of states within each clinical type population.

Keywords: cytokine proteomic biomarkers; computational model; lung disease.

1 Introduction

According to the American Lung Association, an estimated 158,080 Americans are expected to die from lung cancer in 2016 [1]. The 7 lung diseases analyzed here account for some of the most frequent forms of lung disease, with COPD as the fourth leading cause of death in the United States [2]. Respiratory diseases are of multiple origin, and the selected clinical types cover a wide spectrum of suspected causes. More accurate and cost-effective diagnosis is needed so that people with lung diseases are accurately and cost-effectively diagnosed and then treated accordingly, given that Guarascio *et al* declare that not enough is known regarding ideal therapy selection [3].

The use of protein-based biomarkers of lung disease is rapidly advancing, as reviewed by Jun-Chieh *et al* [4], but reliably measuring proteomic biomarker concentrations is difficult due to technical and biological variation, their wide dynamic range of concentrations and numerous post-translational modifications [5]. Despite these variations, we have developed a data-driven Biomarker Computational Model for Lung Disease Classification (BCM-LDC) that reliably distinguishes among various clinically diagnosed lung disease types within our dataset. BCM-LDC hypothesizes that biomarker interactivity induces a distinctive set of host-response protein concentration values for each clinical type, and that certain concentration patterns revealed by these proteins remain characteristically invariant.

BCM-LDC uses a data-driven, supervised-selection learning model; that is, constrained by the limited amount of training data, the model enumerates all possible combinations of biomarker state spaces, then selects that space which most accurately classifies the data into their known clinical types.

In the background §2, we review the suitability of cytokine proteins as host-response biomarkers, the sources of analyzed data, and the difficulties in modeling biological variation given the constraints governing the model, including the issues of overfitting and working within a high-dimensional parameter space. The computational model §3 describes how protein concentrations are topologically modeled and analyzed. §4 presents the experimental results. §5 describes several validation studies, and §6 concludes the paper.

2 Background

2.1 Host-Response Biomarkers

We investigate whether targeted protein variables act as disease state signals due to the existence and modulating strength of their relative and mutual effects upon each other. Our data-driven computational model, BCM-LDC, classifies clinical types and patients directly from the marginal and conditional relationships of biomarker concentration measurements. BCM-LDC selects the unique set of biomarkers – given a small number of biological and statistical assumptions – whose protein host-response topology corresponds to a patient's clinical type. BCM-LDC represents a space of concentration distributions built upon computable discrete states which classifies patients into clinical types, despite significant data variation.

978-1-943436-11-8 / copyright ISCA, BICOB 2018
March 19-21, 2018, Las Vegas, Nevada, USA

Cytokine proteins are secreted by components of the adaptive immune system, and they act as effectors and modulators of lung tissue inflammatory response [6]. The 12 baseline cytokine biomarkers used in this study {EGF, IFNG, IL1A, IL1B, IL2, IL4, IL6, IL8, IL10, MCP1, TNFA, VEGF} (EGF: epidermal growth factor; IFNG: interferon γ; IL: interleukin; MCP: monocyte chemo-attractant protein; TNF: tumor necrosis factor; VEGF: vascular endothelial growth factor) were chosen because of their known sensitivity in host-response to various lung diseases [7], so that concentrations of circulating cytokines in blood serum may be associated with lung disease survival [8].

2.2 Data Sources

BCM-LDC is constructed using host-response cytokine biomarker concentration data from 343 patients given to us in standard units of pico-grams 10^{12} grams per milliliter (pg/ml). Any other data sets obtained from the literature – such as Healthy Serum – are standardized to these units. This baseline data set includes 7 clinical types from which the 12 protein biomarkers are measured. The number of patients per clinical type ranges from 24 to 56 (see Table 4). The **Q=12** baseline biomarkers {EGF, IFNG, IL1A, IL1B, IL2, IL4, IL6, IL8, IL10, MCP1, TNFA, VEGF} measured from each patient's blood serum are chosen because of their known or suspected relationship to lung disease. Two specimens are collected from each patient at the same time, and these two specimens are averaged over each biomarker to provide a single biomarker panel of 12 measurements per patient, except in cases of missing data. Each of the 343 patients are expertly diagnosed as belonging to only one of 7 lung-related clinical types $\boldsymbol{C_t, 1 \leq t \leq 7}$ adenocarcinoma, squamous cell carcinoma, never smokers, smokers with chronic obstructive pulmonary disease (COPD), smokers without COPD, acute lung injury, or cystic fibrosis [9]. We then sequestered a random 10% of these baseline data for subsequent model validation, leaving 310 patients to train to model. There are 659 missing biomarker measurements out of a possible $\boldsymbol{310 * 12 = 3720}$ (82.3% complete) for a total of 3061 measured values. Only 39 of the 343 patients (11.4%) have all 12 biomarker measurements, but 85.4% have 9 or more biomarkers. A total of 17.7% biomarker values are missing from the baseline data set. The mode of the measurements per patient panel is 10. The mean is 9.84. No data was interpolated or averaged to fill in missing data.

Standard protein 2-D gel electrophoresis assay techniques are used to consistently collect homogeneous blood serum specimens. The first five data sets are all from the same unpublished set of experiments [Acknowledgement A] conducted at laboratories at the University of Colorado Health Sciences Center (UCHSC). The last two data sets, cystic fibrosis and acute lung injury, are from different experiments although the wet-lab protocols and analytics are performed in the same way as the first five data sets [Acknowledgement B]. To minimize batch effects, both laboratories incorporated a standard sample in each electrophoresis gel which was subsequently subtracted during analysis, and both used the Cy2 channel from each gel to normalize spot intensities and for automated matching between gels. All patients underwent expert pathology review and have been histologically assigned to one and only one clinical type, provided with the original data sets. The small error bars in Figures 2 and 3 below suggest these data were produced precisely and with quantitative accuracy.

There are many more data values than targeted variables, the 12 biomarkers, which avoids the issue of overfitting. We are working directly with precise concentrations of secreted proteins expressed in blood serum. Even though differences have been uncovered in protein expression between normal and diseased tissues that may have specificity for different tumor types [10], tissue extraction is both costly and invasive. We justify our sampling strategy because it is non-invasive, generates a large set of data with quantitative accuracy involving a small number of targeted variables, and works with a homogeneous composition indicative of the entire organism.

During our initial experiments, we found that any method based upon averaging – such as logistic regression, cosine similarity, or the machine learning Classify function in Wolfram Mathematica© v11.1 – did not classify the baseline clinical types with a sufficient degree of accuracy. Therefore, our subsequent work focused on developing a computational model that processed the entire set of individual concentration values and not just population averages.

3 Computational Model Details

BCM-LDC hypothesizes that interactivity among the biomarkers induces a distinctive concentration distribution as conditioned by the relative concentrations of the other biomarkers. A binning algorithm discretizes the concentration values of every combination of paired biomarkers variables into fixed-sized bins that produces a characteristic multidimensional state space for each clinical type. The binning algorithm is designed to produce both occupied and empty discrete bin states, what we call a *discrete topological structure* (DTS). The bin state pattern that best distinguishes among the clinical type populations is computed by an XOR operation on each possible state space, which also extracts the set of distinctive variable bin states for each clinical type. The distinctive bin state space is then used to assign new patients into one population type given a patient's set of biomarker concentration values. BCM-LDC is briefly presented below.

3.1 Formulating the Computational Model

Our goal is to develop a model that represents the conditional relationships of expressed host-response biomarkers. The first problem is to discretize the biomarker concentration values for every clinical type – paired biomarker combination $\mathbf{CB_r}$, producing a set of bin sizes and number of bins ($\mathbf{W_r}$, $\mathbf{N_r}$, $\mathbf{1 \leq r \leq R}$). A $\mathbf{CB_r}$ is defined as the

aggregate concentration data from each of these pairs of biomarkers within each clinical type. Each clinical type has **77** different combinations of pairs of biomarkers, including pairs of the same biomarker. The binning algorithm bin size computation maximizes the number of occupied bins **Ô** (O-hat), separating concentration data values by the highest possible resolution, while minimizing the number of gaps or empty bins **Õ** (O-tilde) where no data values reside. Empty bins are considered non-permissible data states. BCM-LDC computes a different total number of bins (states) $\mathbf{N_r}$, and bin size $\mathbf{W_r}$, for each combination $\mathbf{CB_r}$. The model computes the probability of each concentration data point belonging to its bin within each $\mathbf{CB_r}$ combination.

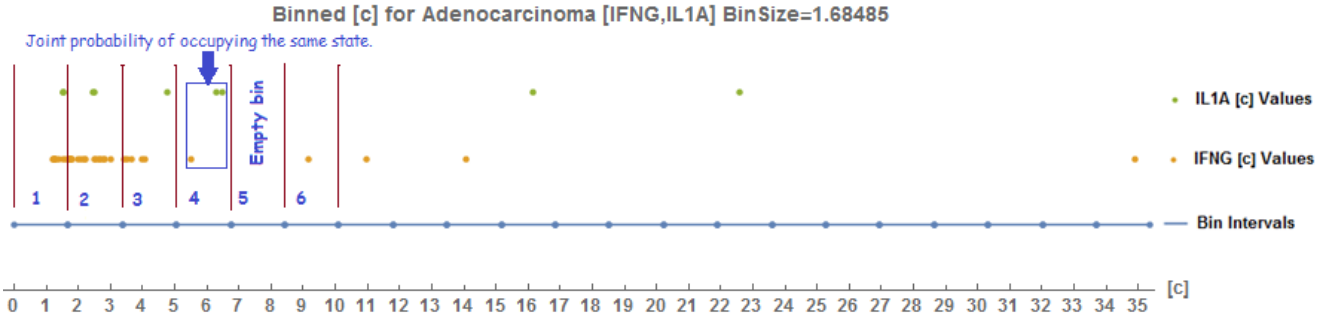

Figure 1: Assigning Adenocarcinoma concentration [c] values to bin states.

3.2 Formulating the Discrete Topological Structure (DTS)

The interactive relationships between each pair of biomarkers {$\mathbf{B_1}$, $\mathbf{B_2}$} are represented by three types of probability. The model computes the pair's *joint occurrence* matrix $\mathbf{M_{C\text{-}joint}}$ – the probability that biomarker $\mathbf{B_2}$ measured at concentration $\mathbf{[c_2]}$ occurs at the same time biomarker $\mathbf{B_1}$ is measured at concentration $\mathbf{[c_1]}$. The model also computes their *conditional probabilities* where, given concentration measurement $\mathbf{[c_1]}$ for $\mathbf{B_1}$, how likely is the measured concentration $\mathbf{[c_2]}$ for $\mathbf{B_2}$. Call this matrix $\mathbf{M_\beta}$. The model uses *marginal probabilities* to represent the influence of individual biomarkers – the probabilities of various concentration values of a subset of biomarker variables without reference to the values of the other variables being considered. Call this matrix $\mathbf{M_\alpha}$. These three types of computed probability taken together express the mutual interactivity and distribution of the biomarker concentration measurements to reveal concentration patterns characteristic of each clinical type. We equate these probability concepts to a *discrete topological structure (DTS)* matrix with equation 1. A data-driven DTS matrix is computed for each $\mathbf{CB_r}$. and the matrix (i.e., the specific set of paired biomarkers) that produces the most accurate set of patient classifications per clinical type is designated $\mathbf{M_C}$ for that population.

$$\mathbf{M_C} = \mathbf{M_{C-joint}}(\mathbf{1} - \mathbf{M_\alpha}) + \mathbf{M_\beta} \qquad (1)$$

In equation 1, $\mathbf{M_{C\text{-}joint}}$ is the population joint occurrence matrix, **1** is a complete matrix of ones (not the identity matrix), $\mathbf{M_\alpha}$ is the **α** interaction matrix of marginal probabilities, and $\mathbf{M_\beta}$ is the **β** interaction matrix of conditional probabilities for the clinical type. The DTS equation is implemented in terms of matrices of conditional and marginal probabilities involving bivariate pairs of biomarkers, each of which are indexed by their respective set of discrete bin states as computed by the binning algorithm. Pseudo-code for the binning algorithm is given in Algorithm 1 below, where $\mathbf{D_r}$ refers to as the combined set of observed concentration data values within each $\mathbf{CB_r}$, for a specific clinical type and biomarker pair {$\mathbf{B_i}$, $\mathbf{B_j}$}.

Algorithm 1: Pseudo-code for the Bin-Min-Max algorithm.

Inputs: D_r: set of concentration data for given CB_r ; maxNbins: max number of bins. **Outputs**: returns W_r, N_r

```
1. foreach combination Dr = { D(Bi), D(Bj) }
2. # Initialize number of bins (Nr), bin step
    size (binInc), bin size (Wr), number of
    empty bins (emptyBins), tmp = 0.
3. Nr ← binInc ← ∜maxNbins
4. Wr ← |max(Dr) - min(Dr)| / Nr
5. emptyBins ← Count_Empty_Bins(Dr, Nr, Wr)
6. result ← |Wr - log_e(emptyBins)|
7. while (result < tmp and Nr < maxNbins -
    binInc) do
8.   Nr ← Nr + binInc
9.   Wr ← |Max(Dr) - Min(Dr)| / Nr
10.  emptyBins ← Count_Empty_Bins(Dr, Nr, Wr)
11.  tmp ← result
12. If (emptyBins > 0) result ← |Wr -
    log_e(emptyBins)|) else result ← 1
13. end while
14. Return (Wr, Nr)
15. end foreach
```

The output of the Max-Bins-Min-Empty-Bins binning algorithm is a bin size $\mathbf{W_r}$ and the number of bins $\mathbf{N_r}$ for each clinical type – paired biomarker combination $\mathbf{CB_r}$. Each value in a set of combined concentration values is assigned to a single bin, but multiple concentration values can be assigned to the same bin, as plotted in Figure 1 for Adenocarcinoma biomarkers {$\mathbf{B_i}$ = IFNG, $\mathbf{B_j}$ = IL1A}. The top 2 **[c]** rows in Figure 1 refer to their actual concentration values measured in pg/ml. These **[c]**values are mapped to specific bin states in the Bin Intervals row. Many of the **[c]** values are grouped in the first few bins. The first 6 states are labeled numerically, and bin 5 is the first empty bin out of

the 23 bins. Bins 1 through 4 illustrate the joint probabilities of IL1A and IFNG values occupying the same state. Additional details for computing each DTS matrix are given in the next section.

3.3 Computing the DTS Matrix $\mathbf{M_C}$

BCM-LDC computes the population joint probabilities for $\mathbf{D_r}$ for each clinical type $\mathbf{C_t}$, combination $\mathbf{CB_r} \in \mathbf{C_t}$, biomarker $\mathbf{B_i} \in \mathbf{CB_r}$, bin $\mathbf{b}$ from 1 to $\mathbf{N_r}$ using equation 2, where $\mathbf{G_b}$ is the number of **[c]** values of $\mathbf{B_i}$ in bin $\mathbf{b}$. The result $\mathbf{P_i}$ is the vector of probabilities for observing the biomarker concentrations in each bin, oftentimes zero. A bin probability equals the number of concentration values $\mathbf{G_b}$ grouped in each bin divided by $|\mathbf{D_r}|$ so that the sum of probabilities over the set of bins is 1.

$$\mathbf{P}_i[\mathbf{b}] = \frac{\mathbf{G_b}}{|\mathbf{D_r}|}, \mathbf{1} \le \mathbf{r} \le \mathbf{R}, \mathbf{1} \le \mathbf{b} \le \mathbf{N_r} \qquad (2)$$

The population joint occurrence matrix $\mathbf{M_{C\text{-}joint}}$ is computed by multiplying each bin probability $\mathbf{P_i}$ for biomarker $\mathbf{B_i}$ with each bin probability $\mathbf{P_j}$ for biomarker $\mathbf{B_j}$, where $\mathbf{B_i}$ is indexed by $\mathbf{i}$ from 1 to the number of bins $\mathbf{N_{Bi}}$ for biomarker $\mathbf{B_i}$ and $\mathbf{B_j}$ is indexed by $\mathbf{j}$ from 1 to the number of bins $\mathbf{N_{Bj}}$ for biomarker $\mathbf{B_j}$. Equation 3 multiplies two vectors (one row vector and one column transposed) together element-wise as an outer product to form a 2-dimensional matrix for that biomarker combination of $\mathbf{B_i}$ and $\mathbf{B_j}$. The dimensions of $\mathbf{M_{C\text{-}joint}}$, one for each $\mathbf{CB_r}$, is $\mathbf{N_{Bi}}$ x $\mathbf{N_{Bj}}$. Bins 1 through 4 in Figure 1 illustrate joint occurrence values greater than zero.

$$\mathbf{M_{C-joint}}(\mathbf{i}, \mathbf{j}) = \mathbf{P}(\mathbf{B_i}) \otimes \mathbf{P}(\mathbf{B_j})^{\mathbf{T}} \qquad (3)$$

The population marginal distributions $\mathbf{M_{i\text{-}marg}}$ and $\mathbf{M_{j\text{-}marg}}$. are computed by equations 4 and 5.

$$\mathbf{M_{i-marg}}(\mathbf{i}) = \sum_{\mathbf{j=1}}^{\mathbf{NBi}} \mathbf{M_{C-joint}}(\mathbf{i}, \mathbf{j}), \mathbf{1} \le \mathbf{i} \le \mathbf{N_{Bi}} \qquad (4)$$

$$\mathbf{M_{j-marg}}(\mathbf{j}) = \sum_{\mathbf{i=1}}^{\mathbf{NBj}} \mathbf{M_{C-joint}}(\mathbf{i}, \mathbf{j}), \mathbf{1} \le \mathbf{j} \le \mathbf{N_{Bj}} \qquad (5)$$

The $\boldsymbol{\alpha}$ interaction matrix $\mathbf{M_\alpha}$ – the matrix from equation 1 with dimensions $\mathbf{N_{Bi}}$ x $\mathbf{N_{Bj}}$ – is composed as the transposition of $\mathbf{M_{i\text{-}marg}}$ repeated $\mathbf{N_{Bj}}$ times. The population conditional probability matrix $\mathbf{M_{C\text{-}cond}}$ for a pair of biomarkers $\{\mathbf{B_i}, \mathbf{B_j}\}$ – one per $\mathbf{CB_r}$ – is computed as an element-by-element matrix division in equation 6.

$$\mathbf{M_{Ci-cond}} = \mathbf{M_{C-joint}}/\mathbf{M_{j-marg}}$$
$$\mathbf{M_{Cj-cond}} = \mathbf{M_{C-joint}}/\mathbf{M_{i-marg}} \qquad (6)$$

The $\boldsymbol{\beta}$ interaction matrix $\mathbf{M_\beta}$ is defined in equation 7 as $\mathbf{P_i}$ divided element-wise by $\mathbf{P_j}$ (from equation 2).

$$\mathbf{M_\beta}(\mathbf{i}, \mathbf{j}) = \mathbf{P_i}/\mathbf{P_j}, \text{given } \mathbf{P_j} > 0, \text{else } 0 \qquad (7)$$

Equation 1, derived from equations 2–7, computes a DTS matrix $\mathbf{M_C}$ for each $\mathbf{CB_r}$ that represents the conditional probability relationship between all pairs of biomarkers within each population. Each $\mathbf{CB_r}$ combination has a characteristic vector of occupied bin states $\hat{\mathbf{O}}$ and empty bin states $\tilde{\mathbf{O}}$ out of a possible number of bins $\mathbf{N_r}$ as calculated by the binning algorithm. Each $\mathbf{CB_r}$ combination now composes an object with the following properties, which will be used to find out the set of distinguishing biomarkers per clinical type:

- clinical type population $\mathbf{C_t}$,
- biomarker pair $\{\mathbf{B_i}, \mathbf{B_j}\}$,
- bin size $\mathbf{W_r}$,
- number of bins $\mathbf{N_r}$,
- bin state vector $[\mathbf{P_i}, \hat{\mathbf{O}}, \tilde{\mathbf{O}}, \mathbf{G_b}]$,
- set of observed concentration values $\mathbf{D_r}$,
- matrices $\mathbf{M_C}$, $\mathbf{M_{C\text{-}cond}}$, $\mathbf{M_{C\text{-}joint}}$, $\mathbf{M_{C\text{-}marg}}$, $\mathbf{M_\alpha}$, $\mathbf{M_\beta}$.

The main advantage of using calculated DTS values instead of raw concentration [c] values is the normalization of scale. Figure 2 plots all biomarker concentration measurements for clinical type Adenocarcinoma, covering a wide range of scales. Figure 3 plots the corresponding Adenocarcinoma DTS values. The binning algorithm calculates the DTS values to all lie within one order of magnitude for every clinical type, and the DTS values are more regularly spaced.

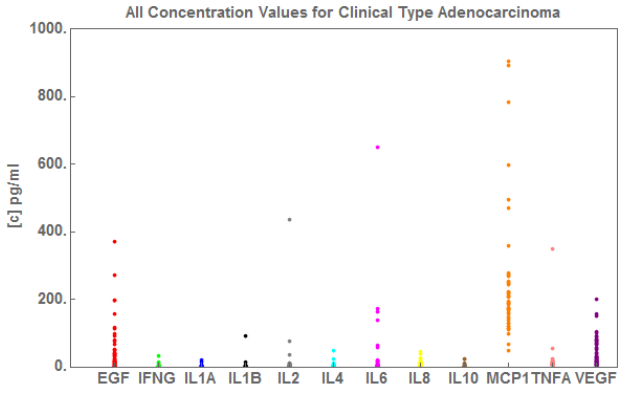

Figure 2: All 12 biomarker Adenocarcinoma [c] values.

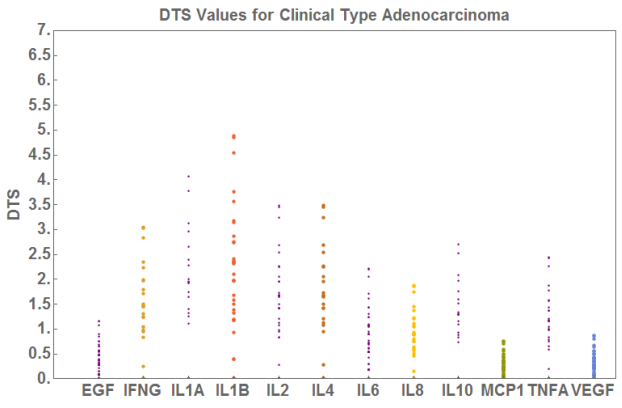

Figure 3: All 12 biomarker Adenocarcinoma DTS values.

3.4 Distinguishing Biomarkers

To reveal the distinguishing biomarkers for each clinical type, BCM-LDC forms a coordinate system of the bin state probability values and the DTS values per biomarker instead of comparing concentration values. The bin states are transformed to matrix form to expose their characteristic and distinguishing states. These integer matrices are constructed by first standardizing the bin state probability and DTS values. The probability values are multiplied by 100 and rounded to integers as percent values along the x-axis to form a standard 100 cells. The corresponding DTS values are raised as exponents to the natural logarithm and rounded to integers, standardizing the y-axis to 256 cells, and starting from the upper left corner. This forms a cellular structure where a whole integer in a cell indicates the presence of a probability–DTS value and 0 otherwise. An element-by-element **XOR** operation between the cellular structures of any two clinical types of the same biomarker reveals which clinical type probability–DTS bin values are unique between those two clinical types. An elaboration of this logic obtains the complete list of distinguishing bin states of the same biomarker among all clinical types. The objective is the same – to identify those matrix cells that are occupied by one and only one clinical type for that biomarker, as described next.

BCM-LDC replaces the occupied matrix integer values with unique $\mathbf{2^n}$ clinical type identifiers (e.g., Adenocarcinoma: 2^1=2), and then adds every matrix together per biomarker so that each matrix cell contains zero, one, or more than one clinical type identifier. An element-by-element $\mathbf{log_2}$ operation that returns a whole integer identifies a single clinical type occupying that cell. This method depends upon the fact that a binomial coefficient (m choose n) (mod 2) is computable using an $_n\mathbf{XOR}_m$ operation. Figure 4 plots the integer matrix for biomarker IFNG for all clinical types, where Adenocarcinoma is distinguished by 3 (red) circled cells. The (blue) circled value of **34**=2+32 indicates that both Adenocarcinoma and Smokers without COPD (2^5=32) exist in the same cell.

0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.
64.	64.	64.	0.	0.	0.	0.	0.	0.	0.	0.	0.
0.	0.	0.	0.	2.	32.	16.	4.	2.	34.	0.	32.
0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.
0.	0.	8.	8.	0.	8.	0.	0.	8.	0.	0.	0.
0.	0.	0.	0.	0.	0.	2	0.	0.	0.	0.	0.
0.	0.	32.	64.	0.	0.	0.	0.	16.	0.	16.	0.
0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.
0.	0.	8.	8.	0.	0.	0.	0.	0.	0.	0.	0.
0.	0.	0.	0.	32.	0.	16.	0.	0.	0.	0.	0.
0.	64.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.
0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.
0.	0.	0.	0.	0.	0.	0.	64.	0.	0.	64.	0.
0.	0.	32.	0.	0.	0.	0.	16.	0.	0.	0.	0.

Figure 4: Partial integer matrix for biomarker IFNG for all 7 clinical types.

The 12 individual integer matrices produced for each clinical type can be consolidated into 3 dimensions to plot their distinguishing biomarkers with respect to the aforementioned probability cell and DTS cell states. Figure 5 plots the distinguishing probability cell and DTS cell states of all the clinical types together. We observe that the range of probability values is low in the Probability dimension – no single biomarker overwhelms any of the others in terms of frequency. It is also clear that the DTS coordinate effectively separates out the clinical types. Interestingly, Never Smokers (blue) displays the most variation among all the clinical types – one is "normal" in a wide variety of states.

4 Experimental Results

Table I lists the common distinguishing biomarkers per clinical type over the 10-fold cross-validation study (see §5.1).

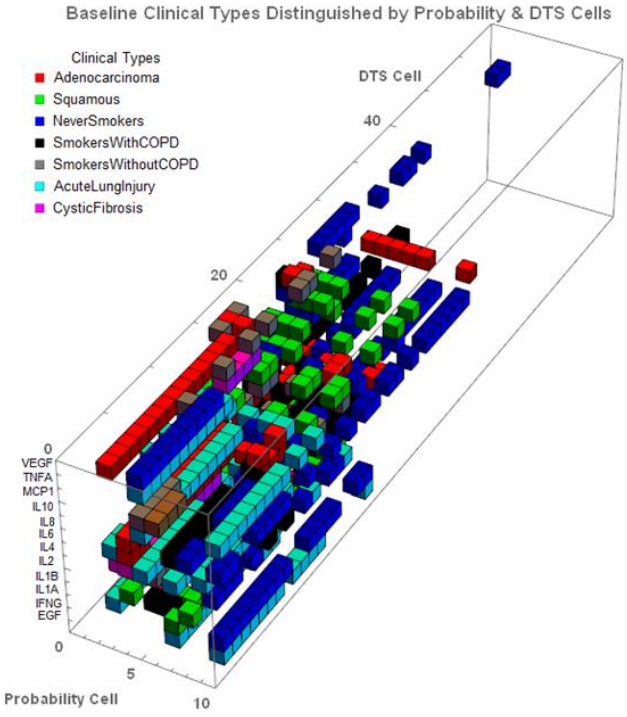

Figure 5: Clinical types distinguished by Probability and DTS states.

TABLE I. DISTINGUISHING BIOMARKERS PER CLINICAL TYPE IN THE PROBABILITY – DTS DIMENSIONS.

Clinical Type Classification C_t	**N Distinguishing Biomarkers**	**Patient Counts**	**Total Bins**	A_t
Adenocarcinoma	**6:** IL1B IL4 IL6 IL8 MCP1 VEGF	53	444	0
Squamous	**6:** IL1B IL2 IL8 IL10 MCP1 TNFA	44	1664	0
Never Smokers	**4:** EGF IFNG TNFA VEGF	55	624	3
Smokers with COPD	**4:** EGF MCP1 TNFA VEGF	49	492	0
Smokers without COPD	**2:** EGF VEGF	53	386	0
Acute Lung Injury	**12:** EGF IFNG IL1A IL1B IL2 IL4 IL6 IL8 IL10 MCP1 TNFA VEGF	62	572	0
Cystic Fibrosis	**1:** IL1A	27	360	0

5 Validation Studies

We can now assign an unknown patient sample **z** to a known clinical type by computing the patient's DTS matrix $\mathbf{M_z}$ and comparing it to every $\mathbf{M_{Ct}}$. Comparing $\mathbf{M_z}$ to every $\mathbf{M_{Ct}}$ uses a fitness function (equation 8) that decides which clinical type is closest to the unknown sample state.

$$\mathbf{s_z} \in \mathbf{C_t} = \mathbf{min}\left(\mathbf{abs}(\mathbf{M_Z} - \mathbf{M_{Ct}})\right), \mathbf{1} \leq \mathbf{t} \leq \mathbf{C_T}|\mathbf{B_j}, \mathbf{N_x} \qquad (8)$$

Assigning a unique bin number and bin probability for each sample biomarker value simply involves looking up the corresponding bin number in the known population probability list for that biomarker. The probability of a sample's concentration value is the expected probability of its assigned bin.

5.1 10-Fold Cross-Validation

We conducted a 10-fold cross-validation study on the 343 baseline patients, where 10% of the samples were randomly extracted 10 times using SQL Server's *NewID* function and then running BCM-LDC over each of the different data partitions. Those distinguishing biomarkers that were present in every one of the 10 runs per clinical type are listed in Table I, column 2. The total number of incorrect baseline patient assignments $\mathbf{A_t}$ over the 10 runs is given in the column 5. Three Never Smokers baseline patients over the 10 runs were incorrectly assigned, of which 2 were the same sample. We account for these incorrect assignments by the large variation present in the Never Smokers patients (see the last part of §3.4) and not by missing biomarker values.

During the same 10-fold cross-validation, each of the 10% sequestered (33) patients were correctly assigned to their respective clinical types with the exception of one (the same) Cystic Fibrosis patient assigned as Acute Lung Injury twice. We account for this incorrect assignment by the tiny sample size of the sequestered Cystic Fibrosis patients, which was the smallest to begin with.

5.2 Healthy Serum Validation

Whereas the baseline clinical types were collected by standard 2-D PAGE gel electrophoresis protocols, measurements from 144 "Healthy Serum" serum samples were taken from a different sampling protocol and experimental design (Luminex® fluorescent bead-based immunoassay [11]). Data was not collected for the EGF or IL2 biomarkers, but included the other 10 biomarkers. When processed along with the baseline data sets, all samples were correctly assigned to their Healthy Serum clinical type.

6 Conclusions

We have developed a computational model, BCM-LDC, that reliably distinguishes among 7 given lung pathologies by assigning biomarker concentration values to discrete states despite significant data variation and technical challenges. BCM-LDC distinguishes the set of biomarker variables that uniquely characterize the clinical types under analysis. The source data – concentration values of host-response serum cytokines – serve as adequate biomarker variables. Excluding Cystic Fibrosis and Smokers without COPD, there is no single biomarker pair that distinguishes among all clinical types, though EGF~VEGF does for 4 types. The minimal biomarker pairs that distinguish among the remaining 5 clinical types are {EGF~TNFA or EGF~VEGF or TNFA~VEGF} and {IL1B~IL8 or IL1B~MCP1 or IL8~MCP1}. Whereas the distinguishing biomarkers extracted are data-driven, patient samples are classified into their single clinical type with reliability greater than 99%.

The Discrete Topological Structure computational model distinguishes among the clinical type populations by discretizing concentrations values to populate only certain bin states. The resulting DTS model simplifies the high-dimensional biomarker concentration space so that some distinguishing features of the lung disease space are revealed.

ACKNOWLEDGMENTS

A. We thank Dr. M. Duncan of USHSC for providing us with 5 / 7 original unpublished data sets.

B. We thank Dr. Paul Bunn of USHSC for providing us with 2 / 7 original unpublished data sets.

REFERENCES

[1] Lung Cancer Fact Sheet from the American Lung Association. Available at http://www.lung.org/lung-health-and-diseases/lung-disease-lookup/lung-cancer/resource-library/lung-cancer-fact-sheet.html (Nov. 2016)

[2] P.T. Reid, J.A. Innes. "Respiratory disease", In: Walker BR, Colledge NR, Ralston SH, Penman ID, eds. Davidson's Principles and Practice of Medicine. 22nd ed. Philadelphia, PA: Elsevier Churchill Livingstone; chap 19. (2014)

[3] A.J. Guarascio, S.M. Ray, C.K. Finch, T.H. Self. "The clinical and economic burden of chronic obstructive pulmonary disease in the USA", ClinicoEconomics and Outcomes Research. Jun 17;5:235-45. (2013)

[4] J.T. Jun-Chieh, C. DeCotiis, A.K. Greenberg, W.N. Rom, "Current Readings: Blood-Based Biomarkers for Lung Cancer", Semin Thorac Cardiovasc Surg. (2013) Winter; 25(4): 328–334.

[5] K. Chandramouli, P-Y Qian, "Proteomics: Challenges, Techniques and Possibilities to Overcome Biological Sample Complexity",Human Genomics and Proteomics, vol 2009, 239204.

[6] R. Sivangala, G. Sumanlatha. "Cytokines that Mediate and Regulate Immune Responses", Austin Publishing Group (2015). Innovative Immunology. Available: www.austinpublishinggroup.com/ebooks

[7] L. Enewold *et al*, "Serum concentrations of cytokines and lung cancer survival in African Americans and Caucasians", Cancer Epidemiol Biomarkers Prev. 2009 Jan;18(1):215-22.

[8] C. A. Dinarello, "Proinflammatory Cytokines", Chest 2000;118;503-508.

[9] J. Subramanian, R. Govindan, "Lung Cancer in Never Smokers: A Review", Journal of Clinical Oncology 25 (5): 561–70. (2007)

[10] M.R. Mehan, D. Ayers *et al*, "Protein Signature of Lung Cancer Tissues", PLoS ONE7(4): e35157. (2012)

[11] Biancotto A, Wank A, Perl S, Cook W, Olnes MJ, *et al.* "Baseline Levels and Temporal Stability of 27 Multiplexed Serum Cytokine Concentrations in Healthy Subjects", PLoS ONE 8(12): e76091. doi:10.1371/journal.pone.0076091 (2013)