Discrete Time Evolution of Proteomic Biomarkers

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

Abstract — We measured a panel of 12 cytokines in seven different populations: i.e., healthy non-smokers, healthy smokers, COPD, Adenocarcinoma and Squamous cell carcinoma of the lung. From these 12 biomarkers of host response to lung disease we have developed a computational and visual model that reliably distinguishes these clinical types. Protein biomarker behavior models are developed as the topological evolution of linear discrete systems from changes in patient protein sample concentrations.

Keywords: proteomics, topological analysis, cytokine biomarker, discrete time evolution.

I. MOTIVATION

Our overall goal is to ensure that people with lung diseases are accurately and cost-effectively diagnosed and then treated accordingly. A significant pair of obstacles in reliably measuring proteomic biomarker concentrations is due to technical and biological variation. We have developed a computational and visual model that reliably distinguishes various clinically diagnosed lung cancer types Our computational model hypothesizes that host biomarker response interactivity induces a distinctive concentration topology and that ensembles of host response proteins are topologically conserved based upon their relative concentration gradients. Whereas patient concentration values vary nearly all the time, certain topological properties remain characteristically invariant according to the patient's clinical type.

The Introduction, §II, describes the suitability of cytokine proteins as host response biomarkers and the sources of data being analyzed. System Characterization, §III, discusses the difficulties in accounting for biological variation and the constraints governing our model, especially the issues of over-fitting and working within a high-dimensional configuration space. §IV on Discrete Time Evolution describes how the stochastic master equation is modified to accommodate our model, particularly in terms of calculating bin states and their joint probability distribution. §V, Topologically Modeling Proteomic Data, introduces computational topology, how protein concentrations can be topologically analyzed, and the computational model that computes topological connectedness in terms of Betti numbers. Experimental Results are given in §VI and Conclusions in §VII.

Acknowledgment: We wish to thank Dr. M. Duncan of USHSC for providing us with these unpublished data sets.

II. INTRODUCTION

A. Cytokine Biomarkers

Cytokines are proteins that are secreted by components of the adaptive immune systems, and they act as effectors or modulators of inflammatory response. Our protein biomarkers – EGF, IFNG, IL1A, IL1B, IL2, IL4, IL6, IL8, IL10, MCP1, TNFA, VEGF (IL: interleukin; EGF: epidermal growth factor; 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 cancers [4], so that concentrations of circulating cytokines in serum may be associated with lung cancer survival [5]. Evidence also suggests that IL6 and IL8 are associated with increased risk of lung cancer [6, 7].

B. Data Sources

Our preliminary model was developed from cytokine measurements of blood samples drawn from patients who have been clinically diagnosed with patients with adenocarcinoma, squamous cell carcinoma, smokers with chronic obstructive pulmonary disease (COPD), as smokers without COPD, or as those who have never smoked [1]. The total number of clinical samples for each of the seven analyzed is 343. The first five data sets are all from the same unpublished set of experiments. The last two data sets are from different experiments, although the wet-lab analytics were performed in the same way by the same lab as the first five data sets. All these data are biomarker host response values given in picograms (10^{-12} gram) per milliliter of individual patients. At least two subsamples were collected from each patient at the same time, and these subsamples were averaged to provide a single sample per patient. Patients were assigned to one and only one clinical type. We are working directly with precise concentrations of secreted proteins expressed in the blood. Our sampling strategy can be justified because it is non-invasive, generated a large set of data with quantitative accuracy involving a small number (12) of targeted variables, and works with a homogeneous composition indicative of the entire organism.

III. SYSTEM CHARACTERIZATION

A. Biological Constraints

Measured protein concentrations in individuals vary across orders of magnitude depending upon the progression of their

978-1-4799-3010-4/14 $31.00 © 2014 IEEE
DOI 10.1109/CSCI.2014.87

disease state, how and where the samples were collected, and by what method and protocol. An overall solution must efficiently navigate within the large and complex probability space in which both the disease and the protein responses occur. Our model is significantly constrained by several conditions, among others.

1) Sample biomarker observations are made of noisy, incomplete and widely variable protein concentration values, so the model(s) and their predictions must be qualitative in nature. The challenge is to make accurate clinical classification and prognostic prediction without relying on statistical inference based upon averaged population concentration values.

2) Protein interaction behavior is qualified using protein concentrations that represent a balance between functional and structural interactions. An interaction produces a change in gradient of either or both of the interacting proteins. The interaction between two proteins depends not only on their binding affinity but also on their concentrations such that the control of protein abundances is an important factor in the functional operation and evolution of natural protein-protein interactions [3]. Even with the wide variation of protein concentrations, their relative abundances may be under tight evolutionary control [2]. These putative structural interactions, involving regions of both high density and very low density, are both important.

3) The range of protein concentrations are also governed by the regulation of kinetic rates. Protein interactions also cover a spectrum of order and function from weakly random to highly-structured because protein function is aggregated from multiple sources. Kumar et al state that "The function of a protein and its properties are decided not only by the static folded three-dimensional structure but also by the distribution and redistributions of the populations of its conformational and dynamic sub-states under different (physical or binding) environments" [10]. These differing levels of organization reveal topological structure.

4) Patients that are sampled are nearly always being clinically treated at the same time, yet the effects of those treatments on relative biomarker concentration levels are usually unknown. This influence is mitigated somewhat by focusing on host response concentrations, but it remains a problem.

B. System Constraints

1) *Over-fitting*. A great deal of proteomics research is plagued by the issue of over-fitting, which occurs when a statistical model describes random error or noise instead of the underlying relationship. Over-fitting generally occurs when a model is excessively complex, such as having too many parameters or experimental variables relative to the number of observations or data points, making it easy to fit multiple models to the data and expose structure that does not correlate with the hypothesis being investigated. Researchers often find that their preliminary or training data fits their model well, but that independent validation study data performs very poorly. We handle this issue by limiting the number of model parameters to only one. Given that proteomics experiments conducted in different laboratories using the same sample-handling protocols can produce drastically different results [8, 9], it is prudent to group and process data sets separately.

2) *High-dimensional configuration space.* Many physical systems can change geometry more easily than they can change topology, and we hypothesize that many interesting proteomic systems can be analyzed as such. For these objects, topological invariants, a map f that assigns the same object to spaces of the same topological type, offer a more meaningful description than linear geometric measures, particularly in terms of configuration space. The notion of configuration space, also called parametric space, is used in molecular biology [11] for representing the space of all possible states of a system characterized by many degrees of freedom. Much of the difficulty in approximating a high-dimensional configuration space is in understanding and simplifying the topology of the space, and little is known about simplifying topology [12]. Yet qualitative equivalence can be determined by looking at a set of configurations in a state space and how a system moves through them. If two systems have the same topological structures in their state spaces, then the two systems are qualitatively the same. We therefore use the topological property of connectedness to suppress the significance of many types of variations and make valid correlations between concentration values and clinical type. The first part to building this computational model is to formulate an equation that navigates within these system constraints.

IV. Discrete Time Evolution

A. Discrete Time Evolution Equation

We start with the differential Stochastic Master Equation as given by van Kampen [14], which describes how the probability of the sample being in a certain state (i.e. a certain set of protein concentrations) changes with time [13] as $t \rightarrow 0$.

$$\frac{\partial}{\partial t}p(X,t) = p(X,t)\left(1 - \sum_{j=1}^{m} \alpha_j \Delta t\right) + \sum_{j=1}^{m} \beta_j \Delta t \qquad (1)$$

In (1), m is the number of possible state interactions in the sample (i.e., those in the study of interest), $\alpha_j \Delta t$ is the probability that interaction j will occur in interval $[t, t + \Delta t]$ given that the system is in state X at time t, and $\beta_j \Delta t$ is the probability that interaction j will bring the system into state X from any other state, say Y. Equation (1) is a gain-loss equation for the probabilities of the separate states. But there are several conceptual issues with this model. First, differential equations presuppose that concentrations of substances vary continuously and deterministically, but these assumptions may be questionable in the genetic regulation of proteins [15, 16]. We also want to treat time t as discrete because of the inherently discrete nature of sampling clinical data. Lastly, the equation assumed that the increments for concentration dc and time dt were common, but in fact they are expressed in different units. The solution to this problem is to represent concentration **[c]** as various state variables under a joint probability distribution and discretize t as step transformations among these various states. This discrete statistical approach places the protein concentration probability distributions into a vector X for each sample as the state variables. A joint probability distribution

$p(X, t)$ then represents the probability that at time t the sample contains X_1 proteins of the first type, X_2 proteins of the second type, and so on. The necessary inputs to the analysis are the distributions of protein concentrations for a clinical sample relative to a standard in pg/µl. The discrete time evolution (DTE) of X in $p(X, t)$ is then understood as the following unit-less quantity.

$$p(X, t + \Delta t) = p(X, t)\left(1 - \sum_{j=1}^{m} \alpha_j \Delta t\right) + \sum_{j=1}^{m} \beta_j \Delta t \qquad (2)$$

$\alpha_j \Delta t$ is computed as the matrix of *marginal* values given state X at time t, and $\beta_j \Delta t$ is computed as the matrix of X / Y at time t. Physically this equation represents a set of discrete states embedded within a continuous range. We adapted the Stochastic Master Equation by computing the necessary state variables, represented by unit-less biomarker concentration probabilities, as discrete bins of concentrations that balances the total number of bins with the number of empty bins.

B. Calculating Bin States

The discrete time evolution of the samples depends upon how the concentration values are binned. We purposely do not exclude any putative outliers because this analysis claims that outliers are legitimate data points due to the topological properties of the ensemble. These bins are biomarker-specific state variables that correspond directly to a concentration binning method that maximizes the number of bins, for the sake of higher data resolution, while minimizing the number of empty bins, for the sake of reducing the number of bin states. This is calculated as the intersection of the curve of the bin size and the curve of the natural log of the number of empty bins, both as a function of the total number of bins. Each bin is then considered as a discrete random variable, a separate state. Each clinical type – marker combination has its own unique (but constant) bin width. Empty bins are considered as non-permissible states. We simultaneously maximize the number of occupied bins for the sake of data point resolution while minimizing the number of unoccupied bins for the sake of reducing the number of possible bin states. The linearity of the ratio of the number of empty bins to the total number of bins for all 7*12=84 clinical type – biomarker combinations allows us to logically compare different clinical types even though their bin widths are numerically different, as shown in Fig. 1. The DTE values for each group are given in Fig. 2-8, and each has a number of distinguishable holes. These holes and boundaries reflect changes in the behavior of the respective proteins and their concentration trajectories.

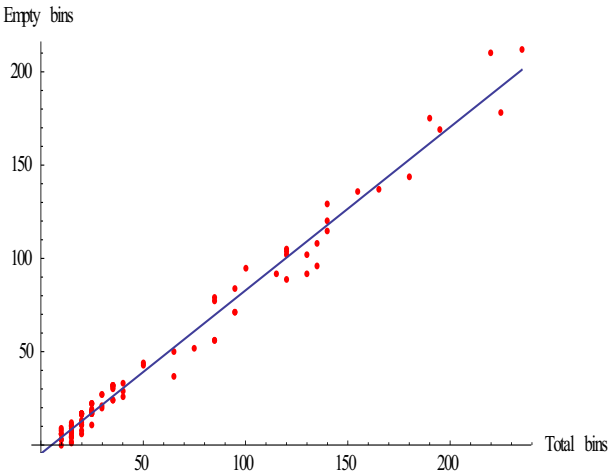

Fig. 1: Ratio of empty bins to total number of bins.

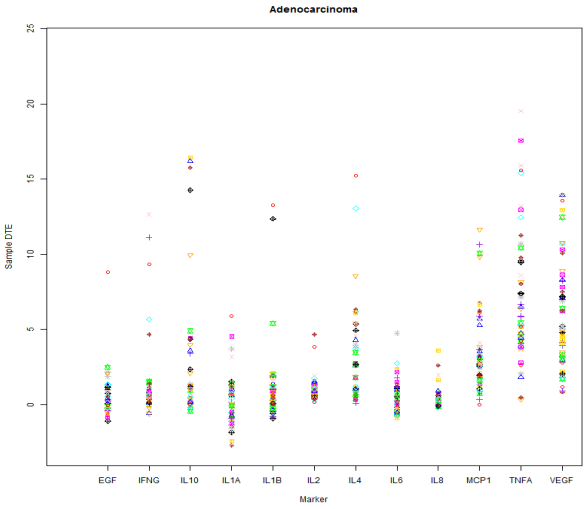

Fig. 2: DTE values for *Adenocarcinoma* group.

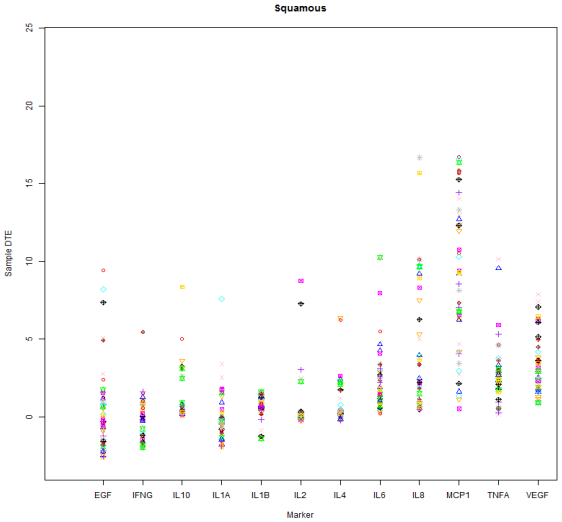

Fig. 3: DTE values for *Squamous* group.

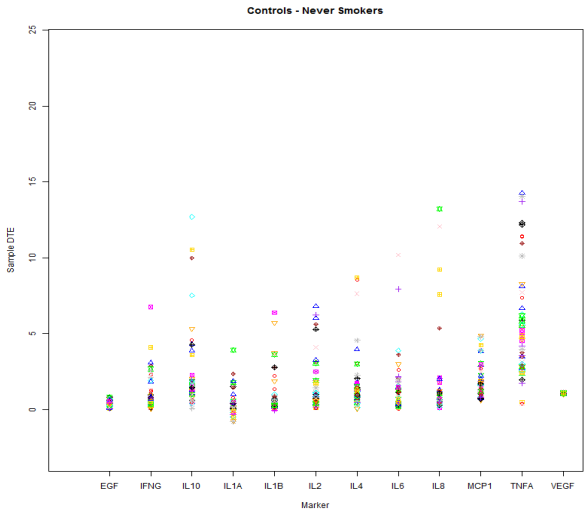

Fig. 4: DTE values for *Never Smokers* group.

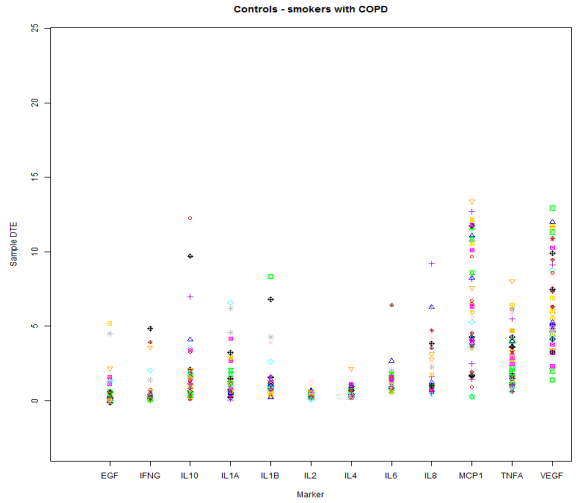

Fig. 5: DTE values for *Smokers with COPD* group.

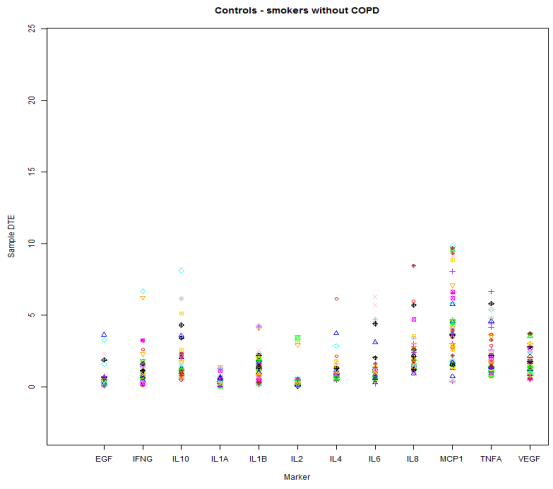

Fig. 6: DTE values for *Smokers without COPD* group.

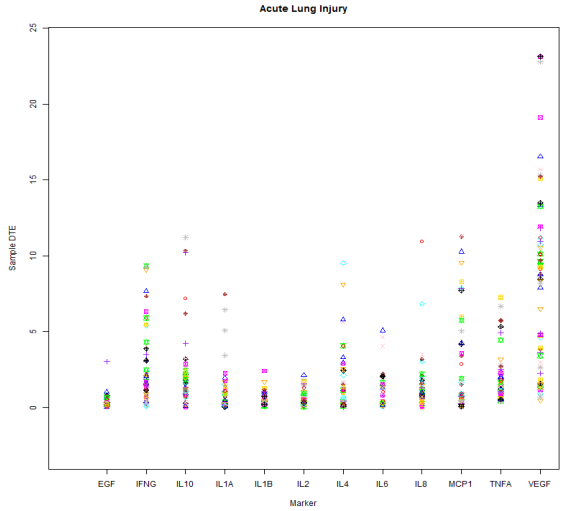

Fig. 7: DTE values for *Acute Lung Injury* group.

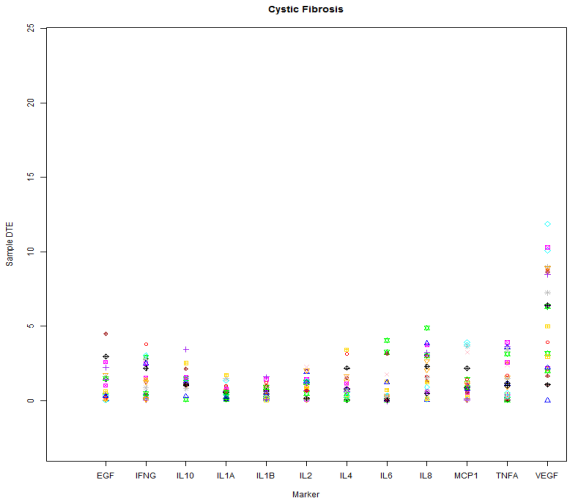

Fig. 8: DTE values for *Cystic Fibrosis* group.

C. Joint Probability Distribution

Calculating the discrete time evolution of the protein biomarker concentrations depends upon how the bin states are interpreted as separate random variables. When random variables are independent, the joint distribution is the product of the marginal densities. If a subset $\boldsymbol{A}$ of the variables $\boldsymbol{X_1}$, …, $\boldsymbol{X_n}$ is conditionally dependent given another subset $\boldsymbol{B}$ of these variables then the joint distribution can be can be efficiently represented by the lower-dimensional probability distributions $\boldsymbol{P(B)}$ and $\boldsymbol{P(A|B)}$. The conditional probability distribution can be calculated by taking the joint density and dividing it by the marginal density of one of the variables. We have incorporated conditional dependencies among the set of biomarker concentration interactions because our implementation of the Stochastic Master Equation directly calculates and incorporates the conditional probability distribution.

V. Topologically Modeling Proteomic Data

A. Computational Topology

A topology is a dynamic system of sets that describes the connectivity of the set. Topological objects can be grouped into classes that have same connectivity. Topological properties include multi-stationary behavior (a stochastic process whose joint probability distribution does not change when shifted in time or space), connectedness (a topological space is said to be connected if it is not the union of two disjoint nonempty open sets), and various feedback mechanisms [17]. The most useful topological invariants involve homology, which defines a sequence of groups describing the "connectedness" of a topological space.

B. Topological Analysis of Proteins

Topology emphasizes those properties of protein systems involving connectivity, continuity, and behavioral space. Many pathological conditions are characterized by pronounced changes in a common set of biochemical variables, at which point a catastrophic response affects how the concentration functionally depends upon the controls. The complexity of cancer biology is exposed by determining those surfaces that reveal the qualitative change from regulated biochemical processes to unregulated.

C. Topological Equivalence and Connectedness

In order to compute connectedness, we must first determine whether samples are topologically equivalent. We compute the Euler characteristic $\chi(s)$ of a point set to distinguish between topologically non-equivalent spaces. To topologically characterize the difference between functional and diseased proteomes then implies that particular clinical type samples reveal topologically non-equivalent concentration surfaces.

Our computational model hypothesizes that statistical ensembles of protein-protein interactions induce a concentration topology, and that certain protein ensembles are topologically conserved based upon their relative concentration gradients. Whereas individual concentration values vary nearly all the time, certain topological properties remain the same under normal biological conditions, and differ substantially under disease conditions. Abnormal concentrations expose a new topological property within a topological space which directly changes the behavior of the respective proteins and their concentration trajectories.

D. Computational Model

We compute topological connectedness as follows.

1) Replace a set of data points with a family of simplicial complexes or simplexes (vertices, edges, triangles, tetrahedra) indexed by a proximity parameter. A simplex is defined as the point set consisting of the convex hull of a set of linear independent points.

2) Analyze these topological complexes using the theory of persistent homology.

3) Encode the persistent homology of a data set in the form of a parameterized version of a Betti number -- a barcode.[18] A barcode can be thought of as the persistence analogue of a Betti number since they represent the data set at various scales.

4) Compute the Betti numbers.

Informally, the kth Betti number refers to the number of unconnected k-dimensional surfaces or the number of k-dimensional holes [19]. A Betti number is the maximum number of cuts that can be made without dividing a surface into two separate pieces. Formally, the nth Betti number is the rank of the nth homology group of a simplicial complex space. The Betti numbers of an object embedded in R3 are respectively:

- β_0 - the number of connected parts separated by gaps,
- β_1 - the number of circles surrounding tunnels,
- β_2 - the number of shells surrounding voids.

Betti intervals describe how the homology of a simplicial complex $X(t)$ changes with t. We want to find Betti intervals that persist for a relatively long time. A filtration or *filter* on a complex X is a collection of sub-complexes $\{X(t) \mid t \leq R\}$ of X such that $X(t) \leq X(s)$ whenever $t \leq s$. A filter basically defines the maximum resolution of the components of the complex. Complex construction is very sensitive to the maximum filtration value, so it is important to establish an algorithm that assigns a consistent filter value.

The algorithm for computing $\beta_0, \beta_1, \beta_2$ proceeds as follows.

1) A point cloud is assigned as the set of points for each biomarker per clinical type. A point cloud is a finite metric space, a finite set of points equipped with a notion of distance.

2) A Euclidean metric space is calculated from the cloud.

3) A *Vietoris-Rips* stream is created using inputs of the maximum dimension (1, 2, or 3), the maximum filtration time, and the number of divisions, which is set to the number of points in the cloud.

4) The number of simplices is calculated from the stream.

5) The persistent intervals are computed using the default simplicial algorithm.

6) The $\beta_0, \beta_1, \beta_2$ Betti numbers are computed.

A Vietoris–Rips complex is an abstract simplicial complex that can be defined from any metric space M and distance δ by forming a simplex for every finite set of points that has diameter at most δ. In a Vietoris-Rips stream, once the filtration value t is greater than the diameter of the point cloud, the Betti numbers will become $\beta_0=1,\beta_1,=\beta_2=$ :::=0. The computed semi-infinite intervals are simply those that persist until $t = t_{max}$. We use the open-source *JavaPlex* library [20] within MATLAB R2013a and 64-bit Java JRE v1.6.32 to compute the Betti numbers.

VI. Topologically Modeling Proteomic Data

A. Point Cloud Data

We model the distance metric of our point clouds as the changing probabilities of the discrete time evolution equation as a function of concentration, where each clinical type is indexed as the log of each of the 12 concentrations, the probability of that concentration happening in the population,

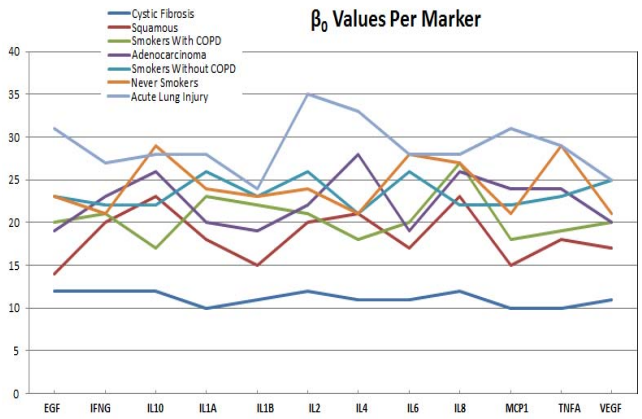

Fig. 9: β_0 values for each clinical type for dimension = 1.

and the computed discrete time evolution. This metric is sufficiently sensitive to distinguish among the clinical types within their first dimension of data, the DTE values.

B. Clinical Type Classification

Fig. 9 plots the results for each clinical type in the first dimension using the filter value that produces a Euler characteristic value of zero, which is computed as the alternating sum of the Betti numbers $\chi(s) = \beta_0 - \beta_1 + \beta_2 - \ldots$ Choosing this filter value allows for a reasonable comparison between the Betti numbers of the various clinical types. Systematically incrementing the filter value before calculating the Betti numbers always produces at least one $\chi(s)$ value that can be found arbitrarily close to 0, in this case computed to 4 decimal places. Monotonically increasing the filter value decreases the $\chi(s)$ value from positive values through zero to negative values.

C. Discussion

We are able to distinguish between all of the clinical types using the computed DTE probabilities and Betti numbers, although it is difficult to assign a physical interpretation to any of the numbers because Betti numbers are not linearly related to each other. Certain markers (IL10, IL1A, IL2, and MCP1) are better at distinguishing among the clinical types because they have no duplicate Betti numbers. Additional validation would use longitudinal sample data to classify subjects as early as possible in their disease progression.

VII. Conclusions

We suppressed much of the variation inherent to using concentration data from proteomic host biomarkers by applying a qualitative topological analysis to reliably distinguish among several lung disease clinical types. The Discrete Time Evolution equation separates the clinical type populations by modeling concentrations as bin states that undergo only certain state transitions. We did not exclude outlier data points and considered empty bin intervals as non-permissible states. The Betti numbers distinguish clinical types, although the list of types cannot be ordered in a linear manner. The Discrete Time Evolution equation served to simplify the high-dimensional biomarker concentration space so that some topology of the lung cancer space was revealed.

References

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

[2] M. Heo, S. Maslov, E. Shakhnovich, "Topology of protein interaction network shapes protein abundances and strengths of their functional and nonspecific interactions," PNAS vol 108 no. 10 4258-4263. (Mar 2011).

[3] J. Zhang, S. Maslov, E. Shakhnovich, "Constraints imposed by non-functional protein-protein interactions on gene expression and proteome size," Molecular Systems Biology 4:210. (2008).

[4] L. Enewold, L. E. Mechanic, E. D. Bowman, Y. L. Zheng, Z. Yu, G. Trivers, A. J. Alberg, C. C. Harris, "Serum concentrations of cytokines and lung cancer survival in African Americans and Caucasians", Cancer Epidemiol Biomarkers Prev. 2009 Jan;18(1):215-22. doi: 10.1158/1055-9965.EPI-08-0705.

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

[6] H. Yanagawa, S. Sone, Y. Takahashi, et al. "Serum levels of interleukin 6 in patients with lung cancer", Br J Cancer. 1995;71(5):1095-1098.

[7] M. Orditura, F. De Vita, G. Catalano, et al, "Elevated serum levels of interleukin-8 in advanced non-small cell lung cancer patients: relationship with prognosis", J Interferon Cytokine Res. 2002;22(11):1129-1135.

[8] Peng, J.; Elias, J. E.; Thoreen, C. C.; Licklider, L. J.; Gygi, S. P. (2003). "Evaluation of multidimensional chromatography coupled with tandem mass spectrometry (LC/LC-MS/MS) for large-scale protein analysis: The yeast proteome". Journal of proteome research 2 (1): 43–50. PubMedId 12643542.

[9] Washburn, M. P.; Wolters, D.; Yates, J. R. (2001). "Large-scale analysis of the yeast proteome by multidimensional protein identification technology". Nature Biotechnology 19 (3): 242–247. PubMedId 11231557.

[10] S. Kumar, B. Ma, C.J. Tsai, N. Sinha, R. Nussinov, "Folding and binding cascades: dynamic landscapes and population shifts," Protein Sci. 9, 10. (2000)

[11] R. Laubenbacher, "System identification of biochemical networks using discrete models," Computation of biochemical pathways and genetic networks, U. Kummer (ed.), Petronius Verlag, Berlin. (2005)

[12] J.M. Kleinberg, "An impossibility theorem for clustering", NIPS 2002: 446-453.

[13] H. de Jong, "Modeling and Simulation of Genetic Regulatory Systems – A Literature Review," J. Comput. Biol. 9 103-129. (2002)

[14] N. G. van Kampen, "Stochastic Processes in Physics and Chemistry", (3rd), Elsevier, Amsterdam. 2007.

[15] D. T. Gillespie, "Exact stochastic simulation of coupled chemical reactions", J. Phys. Chem. 81(25), 2340–2361, 1977.

[16] D. T. Gillespie, "A rigorous derivation of the chemical master equation", Physica D 188, 404–425, 1992.

[17] S. H. Strogatz, "Nonlinear Dynamics and Chaos: With Applications to Physics, Biology, Chemistry, and Engineering," Perseus, New York. (1994)

[18] R. Ghrist, "Barcodes: The Persistent Topology of Data", Bulletin Of The American Mathematical Society, v. 45, Number 1, Jan 2008, pp 61–75.

[19] T. K. Dey, H. Edelsbrunner, S. Guha. "Computational Topology," Advances in Discrete and Computational Geometry, eds. B. Chazelle, J. E. Goodman and R. Pollack. Contemporary Mathematics, AMS, Providence, (1998)

[20] G. Carlsson, "Topology and Data," Bulletin (New Series) Of The American Mathematical Society, Volume 46, Number 2, April 2009, Pages 255–308, Article electronically published on January 29, 2009.