Topological Analysis of Proteomic Data

David Gnabasik and Gita Alaghband
Department of Computer Science and Engineering
University of Colorado Denver
Gita.Alaghband@ucdenver.edu
David.Gnabasik@ucdenver.edu

POSTER PAPER

***Abstract* — Cells become unregulated because topological properties of their interactive behavior make it impossible for them to be regulated. To explore this hypothesis, 2-D PAGE gel electrophoresis protein concentration assay data is embedded and analyzed within a topological computing framework. These 2-D PAGE data are produced using protocols that are standardized, established and widely used. These data are also precise, specific and reliably calibrated alongside an internal reference standard, making the data readily comparable. Importantly, protein concentration data preserve the influence of local interactions, which makes possible a topological analysis. Protein behavior models are developed from longitudinal studies of changes in protein sample concentrations. Continuous variations in protein interactive behavior and their rates of change are shown to produce discontinuous phase shifts in protein concentrations. Preliminary data indicate that a topological analysis of protein data from cancerous cells expose discontinuities in protein behavior space.**

Keywords – proteomics; topological analysis; gel electrophoresis; protein behavior space.

I. INTRODUCTION

Topology is the study of those properties of dynamic systems which remain unchanged under continuous transformations. We seek to determine what information is both sufficient and necessary to produce an information gain from a topological analysis of proteomic data. A topological analysis produces a qualitative information gain for biological systems that involve multiple, parallel local interactions because the global regulation of chemical equilibria involves a large number of simultaneous neighborhood interactions within biological compartments. The existence of local interactions is recognized when a set of different proteins can be introduced into a biological functional container, or space, alongside other chemicals and proteins; transform, bind or interact with and consume or be consumed by these other chemicals and proteins at different rates; and exit the space in a different form all the while the container maintains a specific biologically productive, transformative, or degradation functions. This function is often cast in the form of stable energy minima.

II. SYSTEM INPUTS

The necessary inputs to the analysis are a list of identified protein conformations, the entire spectrum of protein conformations and their modified variants, and their relative concentrations; that is, a set of protein identifiers $\{P_x\}$ and their concentration $\{C_x\}$ relative to a standard in mg/μl at various times T_i. For the topological analysis to recognize discontinuities in protein behavior implies that multiple longitudinal samples also serve as inputs to the analysis. The topological information gain is qualitative because the machinery of chemical rate equations and their kinetic transformations are inferred as optimized topological parameters. The analysis assumes numerous states of equilibrium of stationary, not necessarily minimal, concentration values that expose biochemical signal, rate, and intensity. We measure cellular behavior by examining chemical concentrations.

Multiple chemical and protein concentrations act simultaneously as both sources and products of their interactions. It is essential that simultaneous perturbations or inputs to the system are modeled and simulated both conceptually and computationally. These inputs can arrive in the form of either smooth changes in a particular rate of interaction or smooth changes in relative concentrations. The output space states of this qualitative topological analysis are represented as relative concentrations of the proteins both within each sample and between longitudinal samples. The analysis works in concentration space and seeks to structure its dynamic evolution.

Several significant issues confront a topological analysis of proteomic concentration data. Protein concentrations span multiple orders of magnitude in many biological samples, such as human blood. There are also four broad types of changes in protein concentrations that must be distinguished, those due to:

- random variation due to biological differences, sampling, population characteristics, measurement error, other sources of noise;
- experimental bias due to poor experimental design, sampling and protocol differences, linked or covariant processes;
- overly frequent or non-specific including issues of scale, those proteins that change concentration to a very large number of signal events or stimuli;
- and specific candidate biomarkers, those that are directly relevant to the biological hypotheses; e.g., a specific cancer signal being investigated.

978-1-4673-1382-7/12/$31.00 ©2012 IEEE

It is believed that the best biomarkers constitute a panel of low abundance proteins that are buried under an avalanche of the more abundant proteins [1], which is a problem of determining relative concentrations. How can these variations in concentration be identified and which are the ones to be processed? An efficient way to weed out insignificant variations is to use the topological properties of *connectedness* and disconnectedness. A topological space is said to be connected if it is not the union of two disjoint nonempty open sets. A set is open if it contains no point lying on its boundary. Since continuous variations in protein interactive behavior can produce discontinuous phase shifts, elementary catastrophes in proteomic concentration data sets are identified by a concentration surface that has *folded* where a folded surface represents chemical state.

III. Modeling Topological Proteomics

The goal of topological proteomics is to define the types of topological surfaces of protein interactions and their transformational operators from proteomic data. We treat proteomic data as a dynamic system whose behavior is measured by one set of variables and which is controlled by another set of variables, thereby distinguishing between behavior and control spaces. With 4 dimensions of control there are exactly 7 topologically distinct kinds of discontinuity which can occur in a dynamic system. Many pathological conditions are characterized by pronounced changes in a common set of biochemical parameters at which point a catastrophe happens that affects how the concentration depends upon the controls. The complexity of cancer biology is exposed by finding those folds that cross the line from regulated biochemical processes to unregulated.

A. Representation by Interaction Networks

During any biological process, some proteins act functionally while others act structurally, and it is safe to assume that some serve as both or even switch roles during the process. One would expect that an interaction network would suffice to represent and relate these different functional and structural proteins involved in a particular protein behavior space. Interaction networks possess several advantageous properties. They can represent interaction complexity, incompleteness, noise, scale-free degree distribution, and small-world behavior [2]. Interaction patterns can be used in protein classification to narrow down the scope of hypotheses before experimental verification [3]. However, interaction networks do not easily support a multi-operational system viewpoint, when it is beneficial to view a system in terms of what happens when combinations of operations or functions are performed simultaneously. Multi-operational behavior is represented by specifying interactions between tuples of operations rather than modeling individual operations. In this type of analysis, qualitative heuristics are often needed to effectively model the entire system.

B. Representation by Multi-Operational Tuples

An alternative model for representing relative concentrations is not quantitative but qualitative, where the set of 12 possible interaction states between pairs of proteins A and B are represented as up- ($\uparrow$), down-($\downarrow$), or non-($\bullet$) regulated interactions as shown below.

$\{A \uparrow B \uparrow , A \uparrow B \downarrow , A \uparrow B \bullet\}$ both A and B are up-regulated; A is up-regulated and B is down- regulated; A is up-regulated and B is not-regulated.
$\{B \uparrow A \uparrow , B \uparrow A \downarrow , B \uparrow A \bullet\}$ likewise.
$\{A \downarrow B \uparrow , A \downarrow B \downarrow , A \downarrow B \bullet\}$
$\{B \downarrow A \downarrow , B \downarrow A \uparrow , B \downarrow A \bullet\}$

Space state $A \uparrow B \uparrow$ is distinguished from $B \uparrow A \uparrow$ because perturbations to the system state can be made one protein at a time at a specific point in time. To generalize, we assume that order matters, so that the analysis supports the multi-operational behavior of permutations of protein tuples (e.g., $C \bullet B \downarrow A \uparrow D \uparrow \cdots$). An exhaustive analysis includes the non-($\bullet$) regulated interactions, though it increases the number of possible permutations. A realistic analysis includes only those proteins identified by experiment. The next level of representing relative concentrations is to use integral changes (space states) between tuples of proteins. The final level uses precise concentration measurements for each of these tuples, increasing both the sensitivity of the analysis and its complexity. Computational tractability becomes a product of tuple lengths and the experimentally determined number of space states.

IV. Experimental Results

We analyzed the data from a lung cancer study which sought to identify which cytokine 1 markers (IL-2, IL-4, IL-6, IL-8, IL-10, VEGF, IFN-γ, TNF-α, IL-1α,IL-1β, MCP-1 and EGF) were affected by specific lung diseases (adenocarcinoma, squamous and Chronic Obstructive Pulmonary Disease (COPD). The sample results were analyzed to determine if there were any significant elevations or suppressions in the cytokine 1 markers in relation to each lung disease. 253 patients were involved in this study (gender was not specified) – 52 adenocarcinoma patients, 44 squamous patients, 55 non-smokers, 48 smokers with COPD, and 54 smokers without COPD. Two serum samples were taken from each patient at the same time point and ran in duplication.

References

[1] J. Silberring, P. Ciborowski, "Biomarker discovery and clinical proteomics," Trends Analyt Chem, vol. 29(2): 128, February 2010.

[2] R. Jones, A. Gordus, J. Krall, G. MacBeath, "A quantitative protein interaction network for the ErbB receptors using protein microarrays, " Nature, vol 438: 12, January 2006.

[3] 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. March 8, 2011.