Towards Cell Level Protein Interaction in Multivariate Bioimages

Adnan M. Khan,1, Michael Khan2, Nasir M. Rajpoot1,3,*

1Department of Computer Science, University of Warwick, Coventry CV4 7AL, UK.
2Warwick Medical School, University of Warwick, Coventry CV4 7AL, UK.
3Department of Computer Science and Engineering, Qatar University, Qatar.
*{amkhan, nasir}@dcs.warwick.ac.uk.

Abstract:
Recently, new bioimaging techniques have been proposed to visualize the interaction of several proteins within individual cells. Such techniques could provide the key to understanding complex biological systems such as the protein interactions involved in cancer. Currently, there is a strong urge for sophisticated analytical methods to extract molecular signatures of diseases (such as cancer) in order to not only understand the biological processes behind cancer development but also to aid us in early diagnosis and appropriate treatment. We propose a paradigm for mining of molecular signatures in multivariate bioimages. In contrast to the traditional pixel-level analysis approaches, which ignore cellular structures as units that can be crucial when analyzing cancerous cells, our method incorporates cell-level analysis. We use this framework to perform cell level protein-protein interaction analysis to visualize cellular heterogeneity at a molecular level.

1. Introduction
Bioimage computing is rapidly emerging as a new branch of computational biology which deals with the processing and analysis of bioimages as well as the mining and exploration of useful information present in the vast amounts of image data generated regularly in biology labs around the world. Image based systems biology promises to provide functional localization in space and time. Recent advances in single-molecule detection using fluorescence microscopy imaging technologies allow image analysis to provide access to invisible yet reproducible information extracted from bioimages [1]. Highly multiplexed fluorescence imaging techniques such as Toponome imaging system (TIS) [2] generate massive amounts of multi-channel image data, where each individual channel can provide information about the abundance level of a specific protein molecule localized within an individual cell using the corresponding tag.

In [3], we proposed a framework to identify cell phenotypes present in normal and cancer colon tissues imaged using the TIS microscope. The framework employs non-linear dimensionality reduction (NLDR) and clustering in the embedded space to identify cell phenotypes. In this work, however, we are interested in interaction behavior of protein pairs within each cell phenotype. We want to identify the protein pair interactions which are most relevant in cancer phenotypes and those that are most relevant in normal phenotypes. We hypothesize that the protein pairs that interact a lot might be of interest while defining cell phenotypes within high dimensional space. As the information regarding individual proteins is unavailable in NLDR, due to the derived nature of its coordinates, therefore, in order to extract protein pair interactions, we need a framework that not only identify cell phenotypes and provide visualization but also retain high-dimensional structure for performing such kind of analysis.

2. Cell-level Protein Interaction Framework
The mining framework consists of three stages: preprocessing, feature extraction and clustering. Preprocessing mainly involves alignment of stacks and cell segmentation on DAPI-labeled images to restrict our analysis to cellular areas only. This ensures that signals from stroma and lumen are removed as they can potentially add noise to the subsequent analysis. Features are extracted by computing mean intensity value for each cell across K antibodies (K =
12), which we call the Raw Expression Vector (REV). Let N be the total number of cells found in all the stacks, then the data structure can be represented by an $N \times K$ matrix. Self-organizing map (SOM) [4] is employed to learn cell phenotypes from K-dimensional REV space. SOM is chosen as it is an unsupervised learning algorithm that has the ability to learn abstractions (prototypes) from data in some high-dimensional space without performing dimensionality reduction. It considers each dimension of REV as a component plane - i.e. a slice in the map - and generates a 2-dimensional unified distance matrix (U-Matrix) by combining K component planes in a decision function. Given K component planes, each associated with antibody tag used in this study, we calculate Pearson correlation coefficient for each pair of proteins, localized to an individual cell phenotype, to find out top correlated protein pairs in both cancer and normal samples.

The data used in this study is same as used in [3]. Following table shows top 3 highly correlated protein pairs (in terms of Pearson correlation coefficient) in cancer and normal cells respectively. Most commonly interacting proteins in cancer and normal phenotypes are different and could be the result of changes in protein expression accompanying transformation, adoption of a more stem cell phenotype by some cancer cells. The explanation for similar pairs in normal and cancer (highlighted bold in table) is that there will still be normal or less transformed clones in the cancer sample. The fact that these analyses derive from imaging should allow us to test this hypothesis by looking at the morphology/tissue anatomy. Nevertheless, the framework enables one to easily identify protein pairs which have significantly higher co-expression levels in cancerous tissue samples when compared to normal colon tissue.

<table>
<thead>
<tr>
<th>Correlated</th>
<th>Cancer</th>
<th>Normal</th>
</tr>
</thead>
<tbody>
<tr>
<td>CD166-CD57</td>
<td>Muc2-EpCAM</td>
<td></td>
</tr>
<tr>
<td>CK19-CK20</td>
<td>CD133-CD57</td>
<td></td>
</tr>
<tr>
<td>Muc2-EpCAM</td>
<td>CD166-EpCAM</td>
<td></td>
</tr>
</tbody>
</table>

References
3. A. M. Khan, S.-e.-A. Raza, M. Khan, N. M. Rajpoot, Cell Phenotyping in Multi-Tag Fluorescent Bioimages, Neurocomputing (Submitted).