# A Novel Paradigm for Mining Cell Phenotypes in Multi-Tag Bioimages using a Locality Preserving Nonlinear Embedding

Adnan M. Khan*, Ahmad Humayun†, Shan-e-Ahmad Raza*, Michael Khan∓, Nasir M. Rajpoot*
*Department of Computer Science, University of Warwick, UK
†Georgia Institute of Technology, Atlanta, USA
∓Department of Life Sciences, University of Warwick, UK
Corresponding Authors: {amkhan,nasir}@dcs.warwick.ac.uk

**Abstract.** Multi-tag bioimaging systems such as the toponome imaging system (TIS) require sophisticated analytical methods to extract molecular signatures of various types of cells. In this paper, we present a novel paradigm for mining cell phenotypes based on their high-dimensional co-expression profiles contained within the images generated by the robotically controlled TIS microscope installed at Warwick. The proposed paradigm employs a refined cell segmentation algorithm followed by a locality preserving nonlinear embedding algorithm which is shown to produce significantly better cell classification and phenotype distribution results as compared to its linear counterpart.

**Keywords:** Multivariate fluorescence microscopy, Nonlinear embedding, Cancer biology

## 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 [1]. 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 [2]. Highly multiplexed fluorescence imaging techniques such as MELC or toponome imaging system (TIS) [3] 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. Such high-dimensional representation of multiple co-localized protein expression levels demands for sophisticated analytical methods to extract molecular signatures of diseases such as cancer to not only enable us understand the biological processes behind cancer development but also aid us in early diagnosis and appropriate treatment of cancer. In this paper, we address the problem of mining cell phenotypes based on their high-dimensional protein co-expression

profiles contained within the TIS images generated by a robotically controlled microscope installed at Warwick. We make three important contributions: First, we perform our analysis at the cell level marking a departure from the existing approaches employing pixel-level analysis [3–5]. Second, we show that the raw protein co-expression vectors have a nonlinear high-dimensional structure which can be effectively visualized using a symmetric neighborhood embedding approach. Third, we demonstrate the effectiveness of the nonlinear embedding coordinates for (a) classifying the tissue type at cellular level as compared to principal component analysis (PCA), its linear embedding counterpart, and (b) mining the cell phenotypes in an exploratory clustering setup using affinity propagation [6].

## 2 The Mining Framework

The framework presented in this work consists of three stages: pre-processing involves alignment and cell segmentation, non-linear low-dimensional embedding, and unsupervised clustering.

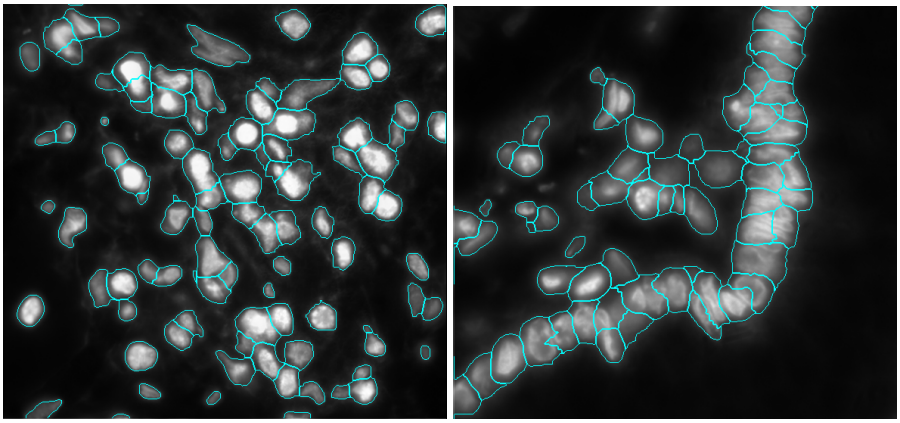
### 2.1 Pre-processing

Raza *et al.* [7] show that the multi-tag images obtained from TIS possess slight mis-alignment, which can potentially introduce noise when finding functional protein complexes in cancerous and normal tissue samples. In line with this argument, the RAMTaB (Robust Alignment of Multi-Tag Bioimages) [7] algorithm is used for aligning multi-tag fluorescent microscopy images.

Cell segmentation is required in order to restrict the analysis to cellular areas only. For nuclei segmentation, we used the multi-step framework proposed by Al-Kofahi *et al.* [8] on DAPI channel which highlights all the nuclei in the image. Initially, an image is binarized using graph-cut based alogorithm to extract foreground. Next, seed points are detected on the foreground of the binarized image by using multiscale Laplacian of Gaussian (LoG) filter, to perform an initial segmentation. Finally, this initial segmentation is refined by using a second graph-cut based algorithm. Nuclei segmentation results obtained using [8] are further post-processed to cater for very small nuclei, often produced as a result of segmentation errors, by either merging with the nearby nuclei or eliminating them altogether (see Figure 1). This further ensures that analysis is restricted to significantly distinguishable cell nuclei only. We refer to this complete process as cell segmentation in the following text.

### 2.2 Raw Expression Vector (REV)

We compute mean intensity value for each cell across $K$ antibodies ($K = 12$) and build a $\boldsymbol{L}_i \in \mathbb{R}^K$ vector for each cell $i$, 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. We normalize the mean intensity value in each column to the range $[0, 1]$ before performing any further analysis.



(*a*) (*b*)

**Fig. 1:** Results of cell segmentation overlaid on top of the two original DAPI images. (*a*) Cancer; (*b*) Normal.

### 2.3 Locality preserving non-linear embedding

Most real-world datasets, regardless of their original dimensionality, contain some structure which should be representable in its intrinsic dimensions. We map all $\boldsymbol{L}_i \in \mathbb{R}^K$ REVs to $\boldsymbol{M}_i \in \mathbb{R}^L$, where $L < K$. t-Distributed Stochastic Neighbor Embedding (t-SNE) [9] is a method which provides such mapping through an optimization aiming to retain the original global and local structure. To achieve this, it define a similarity measure between any two points $i$ and $j$ in the original $\mathbb{R}^K$ space ($p_{j|i}$) and another in the lower dimensional $\mathbb{R}^L$ space ($q_{j|i}$) as,

$$p_{j|i} \propto \exp\left(-\|\boldsymbol{L}_i - \boldsymbol{L}_j\|_2^2 / 2\sigma_i^2\right) \ , \quad q_{j|i} \propto \left(1 + \|\boldsymbol{M}_i - \boldsymbol{M}_j\|_2^2\right)^{-1} \ , \tag{1}$$

where $\sigma_i^2$ is the variance of the Gaussian centered on $\boldsymbol{L}_i$, and $\|\cdot\|_2$ is the Euclidean norm. The user can control this variance, in turn specifying the number of neighbors affecting $p_{j|i}$. In order to keep the inherent structure of the data, t-SNE constrains the similarity measures for any two points to be roughly equivalent between the high and low dimensional space i.e. $p_{j|i} \approx q_{j|i}$. The Kullback-Leibler divergence is a natural fit to impose this contraint. Hence, the cost function to optimize is $\sum_i \sum_j \mathrm{KL}(p_{j|i} \| q_{j|i})$. To make the cost symmetric, all similarity measures are replaced by,

$$p_{j|i} \xrightarrow[\text{replace by}]{} p_{i,j} = \left(p_{j|i} + p_{i|j}\right) / 2N \ . \tag{2}$$

### 2.4 Clustering

Using either the original ($\{\boldsymbol{L}_i\}$) or the lower dimensional data ($\{\boldsymbol{M}_i\}$) we would like to observe the different phenotypes in both cancerous and normal tissue samples. Since each dimension in REV encodes the difference in expression levels

after adding a particular anti-body, it can be used to cluster pixels based on responses to $K$ anti-bodies. To observe the discrimination between cancerous and normal tissue responses, we experimented with two unsupervised clustering methods briefly described below:

**Affinity Propagation Clustering (APC):** APC [6] is an approach where each data samples elects another data point within the dataset to act as its representative or exemplar. The points electing a common exemplar form a single cluster. We initialize the method in a way where each data point has equal likelihood of becoming an exemplar and the final number of clusters is small. The algorithm takes affinity measures between any two points in the dataset as input, which is used in each iteration to find which data points are good exemplars for what samples. To achieve this goal, two kinds of messages are shared between every pair of data points $i$ and $j$. The *responsibility* $r(i,k)$ reflects the suitability point $k$ to represent point $i$ as its exemplar. The reverse message, *availability* $a(i,k)$ defines how much point $k$ thinks it is suited to act as the exemplar of point $i$. Each iteration updates $r(i,k)$ and $a(i,k)$ in a data-driven fashion.

The affinity measure we use between points $i$ and $j$ is

$$K(i,j) = \exp\left(-\|\boldsymbol{M}_i - \boldsymbol{M}_j\|_2^2 / 2\sigma^2\right) , \tag{3}$$

where $\sigma = max(\|\boldsymbol{M}_i - \boldsymbol{M}_j\|_2)/3$. We denote the number of clusters resulting from this approach by $\hat{C}$. It is worth noting that the APC algorithm determines $\hat{C}$ in an unsupervised manner.

**Agglomerative Hierarchical Clustering (AHC):** This is a bottom-up clustering method [10], which starts with each of the $N$ REVs as being a single cluster, and merges two clusters in each iteration. This process can be better represented as a dendrogram tree structure, where cutting across the tree at level $k$ would give $N-k$ clusters [1]. We aim to get the same number of clusters returned by APC, hence we cut the tree at level $\hat{k} = N - \hat{C}$.

The criterion we employ to select the two clusters to merge aims to minimize the increase in the variance of clusters [11]. Mathematically, at each iteration level $k^*$ we have clusters $S_j = \left\{\boldsymbol{M}_{(j,1)}, \ldots, \boldsymbol{M}_{(j,n_j)}\right\}$ where $n_j = |S_j|$ and $j \in \{1, \ldots N-k^*\}$. To make clusters for level $k^*+1$, we seek clusters $\hat{u}$ and $\hat{v}$ such that

$$\hat{u}, \hat{v} = \underset{u,v \in \{1,\ldots N-k^*\}}{\arg\min} \; n_u n_v \left(\|\bar{S}_u - \bar{S}_v\|_2\right)^2 / \left(n_u + n_v\right) , \tag{4}$$

where $\bar{S}_j$ is the mean vector for set $S_j$. This step will result in a new cluster formed by merging $S_{\hat{u}}$ and $S_{\hat{v}}$, hence reducing the number of clusters by 1.

## 3 Experimental Results and Discussion

The data used in this study consists of 3 colon tissue samples. 2 out of the 3 selected tissue samples are taken from healthy colon tissues while 1 taken from

[1] The algorithm starts at level $k=0$, where there are $N$ clusters. Cutting the tree at level $k$ means truncating the tree *after* level $k$

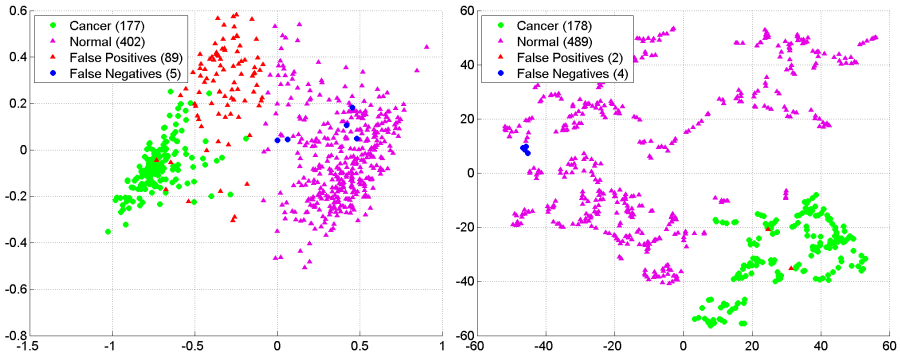



**Fig. 2:** Illustration of low dimensional embedding of Raw Expression Vectors belonging to Cancer and Normal Cells using two different dimensionality reduction techniques: (*a*) PCA; (*b*) t-SNE.

cancerous colon tissue. The tissue samples are verified to be normal or cancerous by independent expert pathologists. A library of 26 antibody tags is used on all 3 colon tissues to generate 3 stacks of multi-tag microscopic bioimages (each stack having 26 tags) using TIS [3]. The library of antibody tags used in this study comprise mainly of nuclei, stem-cell and tumor markers as reported in [12]. Out of these 26 antibodies, some antibodies are ignored because their function is not reflective of the cell activity, while others are discarded because the of the poor quality of its image. Subsequently, only 12 antibody tags are used in the subsequent analysis: CD36,CD44, CD57, CD133, CD166, CK19, CK20, Cyclin-A, Cyclin-D1, CEA, Muc2, EpCAM. Each image is of size $1056 \times 1026$ with pixel resolution of $206 \times 206$ nm/pixel.

All 3 fluorescent microscopy image stacks in the dataset are processed in a similar manner, with image alignment and cell segmentation as described in section 2.1, and finally REV generation as described in section 2.2. For image registration, the default parameters as detailed in [7] are used. For cell segmentation, we tuned for parameters of algorithm in [8] to suit our imaging conditions.

**Experiment 1: Linear (PCA) vs. Non-linear (t-SNE) Dimensionality Reduction for Cell Classification:** For dimensionality reduction, we used two frameworks: one linear (PCA) and other non-linear (t-SNE). Here we show that PCA fails to preserve pairwise relationship between REVs in high dimensional space, whereas t-SNE not only preserves the pairwise relationships but also provides a much superior visual representation of the protein expression vectors. REVs are reduced to 3 dimensions using PCA and t-SNE and *k-means* clustering (with $k = 2$) is applied on these low-dimensional representations to yield the results. Figure 2 shows the visual comparison of clustering results. The results obtained above are evaluated on three quantitative accuracy measures: Sensitivity (Sen), Specificity (Spec), and Positive Predictive Value (PPV). Let $TP$ denotes the number of true positive(cancer cells correctly classified as can-

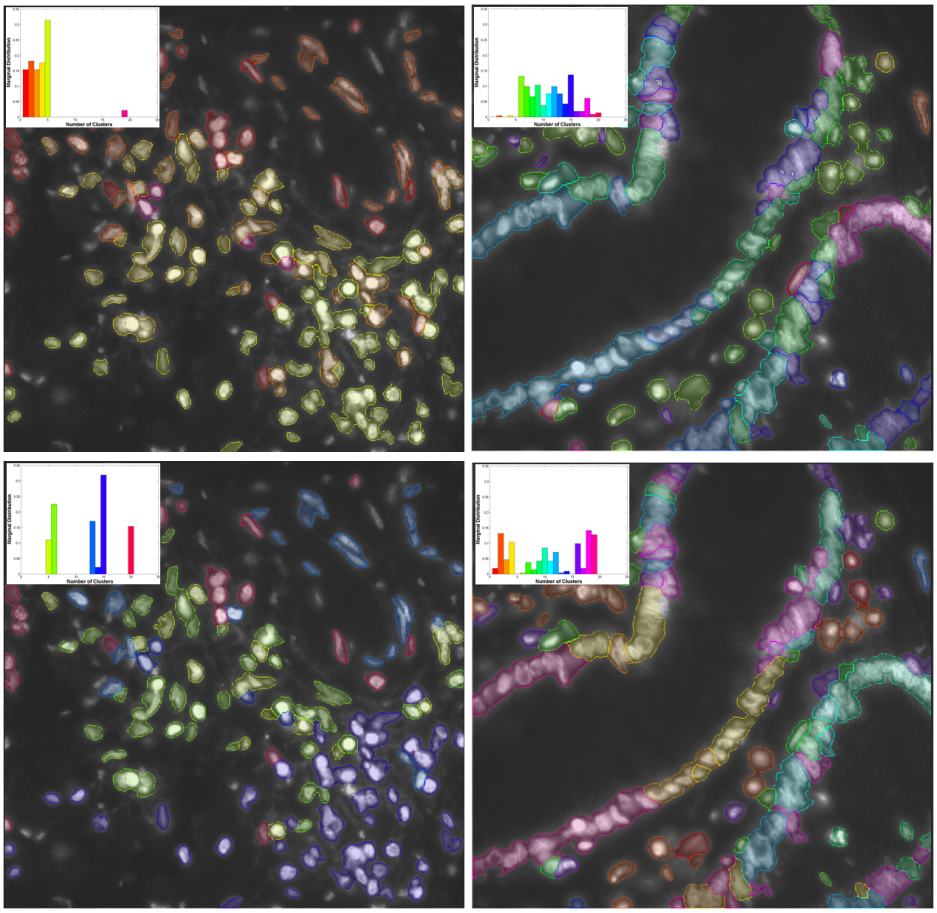


**Fig. 3:** Visual Overlay of 20 phenotypes found using APC(first row) and AHC(second row) on top of DAPI images for a Caner(first column) and Normal(second column) tissue sample. Marginal distributions of cell phenotypes are shown on the top-left corner of each image. Note the difference in distribution of phenotypes in cancer and normal sample.

cerous ), $FP$ the number of false positive (normal cells incorrectly classified as cancerous), $TN$ the number of true negatives (cancer cells incorrectly classified as normal), and $FN$ the number of false negatives (normal cells correctly classified as normal), then Sen is defined as $TP/(TP+FN)$, Spec as $TN/(TN+FP)$ and PPV as $TP/(TP+FP)$. Table 1 shows quantitative comparison of classification (using *k-means*, with $k = 2$) when (1) 12−dimensional REVs; (2) 3−dimensional PCA; and (3) 3−dimensional t-SNE data are used for classification. Note that the PPV for t-SNE is approximately 32% higher than those of the original data and PCA.

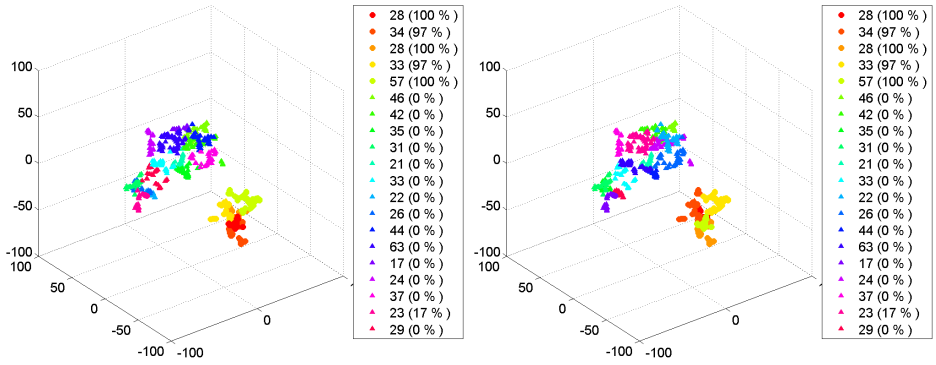



**Fig. 4:** Scatter plot of t-SNE embedding of REVs, where cancer (circles) and normal cells (triangles) are colored using 20 different phenotypes identified using APC(left) and AHC(right). The legend displays the number of cells present in the cluster and percentage of cancerous cells respectively, identified in the corresponding phenotype. Note that phenotypes are marked on the basis of majority; i.e. a given phenotype is marked as cancer (circle) if majority of its cells belong to cancer and viceversa.

**Experiment 2: Cell Phenotype Analysis using Unsupervised Clustering:** Given the promising cell classification results obtained above, we are further interested in finding different phenotypes present in normal and cancer samples. Unsupervised clustering can be used for such type of analysis. Here, we present a comparative analysis of different phenotypes identified by using two popular clustering frameworks described in section 2.4: APC and AHC. Number of clusters ($\hat{C}$) is identified using APC and the same number is used in AHC.

Figure 3 shows the overlay of different cell phenotypes identified using APC and AHC over DAPI Images, whereas Figure 4 shows the scatter plot of different phenotypes. In order to quantitatively assess the performance of all cell phenotype mining methods used here, we employ the average of symmetric KL-divergence of cell phenotype distributions between cancer/normal and normal/normal samples. Results shown in Table 2 again demonstrates the effectiveness of t-SNE as compared to PCA.

## 4 Conclusions

We presented a paradigm for cell-level mining of molecular signatures in multi-tag bioimages using a nonlinear embedding approach. This approach is a marked departure from the traditional pixel-level approaches. We showed that the symmetric neighborhood embedding outperforms the original high-dimensional raw protein expression vectors in terms of its ability to discriminate between normal and cancer tissue samples on the basis of their phenotypic distributions. Our future work will employ this paradigm in a large-scale validation for extracting biologically plausible molecular signatures of various cell phenotypes found in cancer specimens.

| | Sen | Spec | PPV |
|---|---|---|---|
| **REV** | 0.9725 | 0.8187 | 0.6654 |
| **PCA** | 0.9725 | 0.8187 | 0.6654 |
| **t-SNE** | 0.9780 | 0.9959 | **0.9889** |

**Table 1:** Quantitative Comparison of Classification Results using: 12−dimensional REV; 3−dimensional PCA; 3−dimensional t-SNE. Values marked in bold show best results. On all 3 scales, t-SNE embedding gives superior performance.

| | Cluster | Inter-class | Intra-class |
|---|---|---|---|
| **REV** | AP | 29.8565 | 2.3992 |
| | HC | 33.7972 | 2.9332 |
| **PCA** | AP | 22.9484 | **0.2841** |
| | HC | 26.0002 | 0.697 |
| **t-SNE** | AP | **41.896** | 0.6915 |
| | HC | 41.7355 | 0.8345 |

**Table 2:** Quantitative comparison of inter- and intra-class symmetric KL-divergence of cell phenotype distributions for 12−dimensional REV, 3−dimensional PCA, and 3−dimensional t-SNE using APC and AHC.

## References


1. Megason, S., Fraser, S.: Imaging in systems biology. Cell **130**(5) (2007) 784–795
2. Danuser, G.: Computer vision in cell biology. Cell **147**(5) (2011) 973–978
3. Schubert, W., Bonnekoh, B., Pommer, A., Philipsen, L., Böckelmann, R., Malykh, Y., Gollnick, H., Friedenberger, M., Bode, M., Dress, A.: Analyzing proteome topology and function by automated multidimensional fluorescence microscopy. Nature biotechnology **24**(10) (2006) 1270–1278
4. Humayun, A., Raza, S.e.A., Waddington, C., Abouna, S., Khan, M., Rajpoot, N.M.: A Framework for Molecular Co-Expression Pattern Analysis in Multi-Channel Toponome Fluorescence Images. In: Microscopy Image Analysis with Applications in Biology (MIAAB). (Sept. 2011)
5. Kölling, J., Langenkämper, D., Abouna, S., Khan, M., Nattkemper, T.: Whide-a web tool for visual data mining colocation patterns in multivariate bioimages. Bioinformatics (2012)
6. Frey, B.J., Dueck, D.: Clustering by passing messages between data points. Science **315**(5814) (2007) 972–976
7. Raza, S.e.A., Humayun, A., Abouna, S., Nattkemper, T.W., Epstein, D.B.A., Khan, M., Rajpoot, N.M.: Ramtab: Robust alignment of multi-tag bioimages. PLoS ONE **7**(2) (2012)
8. Al-Kofahi, Y., Lassoued, W., Lee, W., Roysam, B.: Improved automatic detection and segmentation of cell nuclei in histopathology images. Biomedical Engineering, IEEE Transactions on **57**(4) (2010) 841–852
9. Van der Maaten, L., Hinton, G.: Visualizing data using t-sne. Journal of Machine Learning Research **9**(2579-2605) (2008) 85
10. Jain, A., Murty, M., Flynn, P.: Data clustering: a review. ACM computing surveys (CSUR) **31**(3) (1999) 264–323
11. Ward Jr, J.: Hierarchical grouping to optimize an objective function. Journal of the American statistical association (1963) 236–244
12. Bhattacharya, S., Mathew, G., Ruban, E., Epstein, D., Krusche, A., Hillert, R., Schubert, W., Khan, M.: Toponome imaging system: in situ protein network mapping in normal and cancerous colon from the same patient reveals more than five-thousand cancer specific protein clusters and their sub-cellular annotation by using a three symbol code. Journal of Proteome Research (2010)