BIOINFORMATICS

Vol. 00 no. 00 2013
Pages 1–7

DiSWOP: A Novel Measure for Cell-Level Protein Network Analysis in Localised Proteomics Image Data

Violeta N. Kovacheva [1,*], Adnan M. Khan [2], David Epstein [3], Michael Khan [4] and Nasir M. Rajpoot [2,5,*]

[1]Department of Systems Biology, The University of Warwick, Coventry CV4 7AL, UK
[2]Department of Computer Science, The University of Warwick, Coventry CV4 7AL, UK
[3]Mathematics Institute, The University of Warwick, Coventry CV4 7AL, UK
[4]School of Life Science, The University of Warwick, Coventry CV4 7AL, UK
[5]Department of Computer Science and Engineering, Qatar University, Qatar

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT

Motivation: New bioimaging techniques have recently been proposed to visualise the colocation or interaction of several proteins within individual cells, displaying the heterogeneity of neighbouring cells within the same tissue specimen. Such techniques could hold the key to understanding complex biological systems such as the protein interactions involved in cancer. However, there is a need for new algorithmic approaches that analyse the large amounts of multi-tag bioimage data (also known as localised proteomic or toponomic data) from cancerous and normal tissue specimens in order to begin to infer protein networks and unravel the cellular heterogeneity at a molecular level.

Results: The proposed approach analyses cell phenotypes in normal and cancerous colon tissue imaged using the robotically controlled Toponome Imaging System (TIS) microscope. It involves segmenting the DAPI-labelled image into cells and determining the cell phenotypes according to their protein-protein co-dependence profile. These were analysed using two new measures, Difference in Sums of Weighted cO-dependence/Anti-co-dependence profiles (DiSWOP and DiSWAP) for overall co-expression and anti-co-expression, respectively. These novel quantities were extracted using 11 TIS image stacks from either cancerous and normal human colorectal specimens. This approach enables one to easily identify protein pairs which have significantly higher/lower co-expression levels in cancerous tissue samples when compared to normal colon tissue.

Availability: http://www2.warwick.ac.uk/fac/sci/dcs/research/combi/research/bic/diswop

Contact: v.n.kovacheva; n.m.rajpoot@warwick.ac.uk

1 INTRODUCTION

In order to understand cellular biology on a systems level, relationships between molecular components must be understood not only at a functional level but also localised in the spatial domain [Megason and Fraser, 2007]. This is due to the fact that proximity of key proteins provides an indication of the possible existence of functional protein complexes. As a consequence, new bioimaging techniques have been recently proposed to visualise the colocation or interaction of several proteins in cells in intact tissue specimen. These include MALDI imaging [Cornett *et al*., 2007], Raman microscopy [van Manen *et al*., 2005], Toponome Imaging System (TIS) [Schubert *et al*., 2006] and multi-spectral imaging methods [Barash *et al*., 2010]. TIS is an automated high-throughput technique able to co-map up to a hundred different proteins or other tag-recognisable bio-molecules in the same pixel on a single tissue section [Schubert *et al*., 2012]. It runs cycles of fluorescence tagging, imaging and soft bleaching *in situ*. While colocation does not necessarily imply interaction, it has been consistently found that clusters containing particular proteins are found in specific sub-cellular compartments, hence allowing such a hypothesis to be generated [Bhattacharya *et al*., 2010]. Also, a frequently occurring colocalisation of proteins indicates a possible functional physiochemical interaction. TIS has a sub-cellular maximum lateral resolution of 206×206 nm/pixel [Kolling *et al*., 2012] which allows the determination of sub-cellular protein network architectures and, therefore, can potentially reveal the sub-cellular toponome. The combination of proteomic information with spatial sub-cellular level topographical data has been termed 'toponomics' [Schubert *et al*., 2003, 2012]. We can advance our understanding of the toponome by finding correlations between phenotype, function, and morphology of cells.

Biomarkers used in current clinical practice are limited to the simultaneous analysis of only a handful of proteins. They, therefore, fail to assess the true complexity of cancer, and the resulting biomarkers have a low prognostic value [Vucic *et al*., 2012]. The capabilities of the TIS hold promise for developing a new generation of multiplex biomarkers [Evans *et al*., 2012] which could aid the development of personalised medicine. Studying the cancer toponome could uncover previously unknown mechanisms of tumour formation and could identify new potential drug targets in the form of protein interactions.

[Bhattacharya *et al*., 2010] have shown how TIS imaging can be used in cancer research for protein network mapping. However,

*to whom correspondence should be addressed

there is a need for new algorithmic approaches that analyse the co-expression patterns. The standard way to analyse TIS images is to apply a threshold to each image of the stack and so reduce it to binary values [Schubert *et al.*, 2006]. However, while this step is straight forward, it is bias-prone, subjective and time-consuming. Furthermore, by reducing the image to binary, a lot of potentially very important information could be lost. [Langenkamper *et al.*, 2011] and [Humayun *et al.*, 2011] have both presented such non-threshold methods. Their algorithms cluster molecular co-expression patterns (MCEPs) on a pixel level and therefore lose the variation at a cell level. This can be crucial when analyzing cancerous samples due to the heterogeneity of cancer cells [Vucic *et al.*, 2012]. Furthermore, these algorithms are based on the raw expression levels, which are intensity dependent and hence may vary between different stacks. A similar approach is used in the Web-based Hyperbolic Image Data Explorer (WHIDE) [Kolling *et al.*, 2012], which allows analysis of the space and colocation using a H2SOM clustering [Ontrup and Ritter, 2006]. While this tool is very effective at identifying molecular co-expression patterns, the cellular structure is lost and hence the method is unable to analyse the different cell phenotypes that may be present in the samples.

In this paper, a new approach is proposed where the protein-protein dependence profile (PPDP) is considered instead of the raw protein expression profiles. There has been evidence in the literature that despite the spherical and the exploratory cell states of rhabdomyosarcoma cells having identical average protein profiles, striking differences were found between the two states at the sub-cellular protein cluster level [Schubert, 2010]. Hence, rearrangement, rather than up- or down-regulation of proteins is (or can be) key to generating new cell functionalities [Schubert *et al.*, 2012]. This shows the importance of co-dependence of proteins rather than abundance on its own. Furthermore, we perform the analysis at cell level rather than pixel level, allowing for the cells to be phenotyped according to their PPDP. This enables us to gain a better understanding of the heterogeneity within the cancer cell population. Lastly, two new measures are proposed to enable us to infer small-scale protein networks. These new measures highlight protein pairs which have very different interaction in cancer and normal tissue. An overview of the approach is presented in Figure 1. Applying it to synthetically generated data gave the expected results, giving confidence in the new measures.

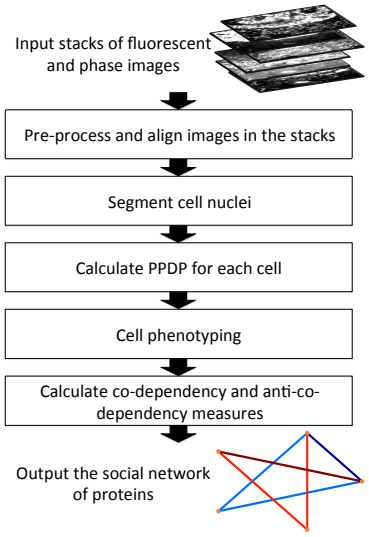

Fig. 1. Overview of the proposed framework.

2 METHODS

2.1 Data and pre-processing

The image data used in this study was acquired using a TIS microscope [Schubert *et al.*, 2006] installed at the University of Warwick. Samples had been surgically removed from colon cancer patients. One sample was taken from the surface of the tumour mass, and another one was selected from apparently healthy colonic mucosa at least 10cm away from the visible margin of the tumour. Two visual fields were manually selected in each tissue sample, resulting in four TIS data sets from a single patient. The results presented here were obtained by considering a total of 11 samples – 6 healthy and 5 cancerous. A library of 26 antibody tags, some of which are known tumour markers or cancer stem cell markers, were used based on the findings by [Bhattacharya *et al.*, 2010]. CD133, CK19, Cyclin A, Muc2, CEA, CD166, CD36, CD44, CD57, CK20, Cyclin D1 and EpCAM were used in the analysis with the rest being excluded either because their function was not related to the cell activity, e.g. DAPI localises the cell nuclei, or because the image of their expression in one or more of the samples was of a poor quality.

Background autofluorescence is digitally subtracted at an early stage. Hence, any remaining fluorescence should be true protein expression. In each of the stacks, the images were aligned using the RAMTaB (Robust Alignment of Multi-Tag Bioimages) algorithm [Raza *et al.*, 2012]. This is done in order to prevent possible noise resulting from the slight mis-alignment of the multi-tag images obtained using TIS. Then, if there are K tags, each having a corresponding image of size m by n, the data can be represented as a $K \times mn$ matrix

$$\mathbf{X} = \left[\begin{array}{c|c|c|c} x^1_{1,1} & x^1_{1,2} & \cdots & x^1_{m,n} \\ x^2_{1,1} & x^2_{1,2} & \cdots & x^2_{m,n} \\ \vdots & \vdots & \ddots & \vdots \\ x^K_{1,1} & x^K_{1,2} & \cdots & x^K_{m,n} \end{array}\right], \quad (1)$$

where $x^k_{i,j}$ is the expression level of protein k at pixel (i, j).

In [Khan *et al.*, 2012], we proposed to perform cell segmentation of TIS stacks in order to restrict the 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. To follow best practice, one should segment entire cells since some of the proteins observed are located in parts of the cells other than the nucleus, such as the cytoplasm, vesicles or the Golgi apparatus. However, this is challenging in cancerous tissues because of the variable orientation of cells due to disrupted tissue architecture and a tag of the cell membrane was not available to us to precisely identify entire cells. Instead, each image was segmented using a modified form of the graph cut method [Al-Kofahi *et al.*, 2010] proposed by our group [Khan *et al.*, 2012] applied to a DAPI channel. This was necessary in order to extract pixel locations of the nuclei and their immediate neighbourhood only, as the DAPI tag stains the DNA. Using only nuclei may reduce the amount of cell available for analysis but is comparatively unambiguous. Details of the method and examples can be found in the Supplementary Materials.

The cell-localised protein expression values for each of the K proteins is collected in a protein expression matrix $\mathbf{X}_c$ of the order $K \times N_c$ for each cell c

$$\mathbf{X}_c = \{\mathbf{x}_{i,j} | \, (i,j) \in \Omega_c\}, \quad (2)$$

where $\Omega_c = \{(i_1, j_1), (i_2, j_2), ..., (i_{N_c}, j_{N_c})\}$ denotes the set of pixel coordinates in cell c, $N_c = |\Omega_c|$ denotes the number of pixels in each cell c

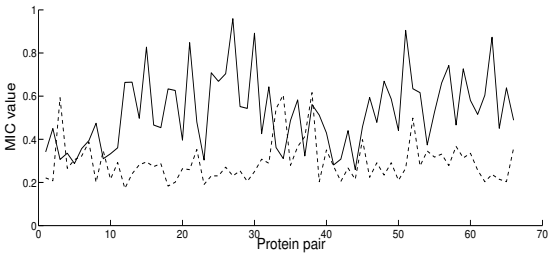

Fig. 2. Protein-protein dependence profile (PPDP) of two cells from the same specimen.

and the vector $\mathbf{x}_{i,j} = [x^1_{i,j}, ..., x^K_{i,j}]$ is the expression levels of each tag at pixel (i, j). In matrix form this is given by

$$\mathbf{X}_c = \begin{bmatrix} x^1_{i_1,j_1} & x^1_{i_2,j_2} & \cdots & x^1_{i_{N_c},j_{N_c}} \\ x^2_{i_1,j_1} & x^2_{i_2,j_2} & \cdots & x^2_{i_{N_c},j_{N_c}} \\ \vdots & \vdots & \ddots & \vdots \\ x^K_{i_1,j_1} & x^K_{i_2,j_2} & \cdots & x^K_{i_{N_c},j_{N_c}} \end{bmatrix}. \quad (3)$$

2.2 Protein-protein dependence profile (PPDP)

The pairwise maximal information coefficient (MIC) [Reshef *et al*., 2011] for each pair of proteins, localised to an individual cell c, is calculated to obtain the protein-protein dependence profile (PPDP) of the cell. We used this statistic since it has been shown to capture a wide range of associations, both functional and not, and it gives similar scores to equally noisy relationships of different types [Reshef *et al*., 2011]. Details of the way it is calculated can be found in the Supplementary Materials. For each cell c, a $K(K-1)/2$-dimensional vector μ_c of pairwise MIC scores is obtained. The vector represents the PPDP of the cell and can be expressed as

$$\mu_c = \left[\mu_c^{1,2}\mu_c^{1,3}\cdots\mu_c^{1,K}\mu_c^{2,3}\mu_c^{2,4}\cdots\mu_c^{2,K}\cdots\mu_c^{K-1,K}\right], \quad (4)$$

where $\mu_c^{i,j} \in [0, 1]$ is given by the MIC between rows i and j of the matrix $\mathbf{X}_c$. The PPDP for two sample cells from the same tissue specimen is shown in Fig 2.

Other co-dependence measures were also considered for the analysis. Pearson's and Spearman correlations fail to capture non-linear relationships between protein expression profiles, which often occur due to the inhomogeneous structure of the cells. Mutual information and normalised mean expression values were also tested. However, each of these resulted in a batching effect where some clusters were predominantly located in a single, usually cancerous, sample (Supplementary Figure 2). This seems biologically unlikely as we expect that there should be some normal cells within the tumour tissue and that cancers share some common types of cells. These findings are consistent with the findings by [Schubert, 2010] that functionality can be determined by colocation rather than changes in abundance levels.

2.3 Cell phenotyping based on localised PPDP

The vector μ_c is the PPDP of the cell c and can be used to determine the cell phenotype using a clustering algorithm. Affinity Propagation (AP) is a clustering method, which takes as input a matrix containing measures of similarity between pairs of data points. Real-valued messages are passed between data points until a high-quality set of exemplars and corresponding set of clusters gradually emerges [Frey and Dueck, 2007]. We have used a Gaussian Kernel based on the Euclidean distance between the protein co-dependence profiles of cells as an affinity matrix, so for a pair of cells a and b with PPDPs μ_a and μ_b, respectively, the (a, b) entry of the similarity matrix (for $a \neq b$) is given by

$$s_{a,b} = exp\left(\frac{-\|\mu_a - \mu_b\|^2}{2\sigma^2}\right), \quad (5)$$

where $\sigma = \left(max_{a,b}\|\mu_a - \mu_b\|\right)/3$ and $\|\cdot\|$ is the Euclidean distance. All diagonal entries of the matrix are set to equal the minimum value of the matrix. This means that each cell is equally likely to be a cluster centroid and results in a moderate number of clusters. We denote the number of cell phenotypes resulting from this approach by $\hat{C}$, which in this instance was found to be 41. The phenotyping results for a normal and a cancer samples are shown in Supplementary Figure 3. An Agglomerative hierarchical clustering approach with the same number of clusters was also considered. It was encouraging to see that the hierarchical clustering gave very similar results but the results have not been shown here.

2.4 Protein-protein co-dependence and anti-co-dependence measures

Once the cell phenotype clusters have been obtained, an average PPDP, $\bar{\mu}_S$ is calculated for each cluster S. For a protein pair (i, j) (with $i < j \leq K$) $\bar{\mu}_S^{i,j}$ is given by

$$\bar{\mu}_S^{i,j} = \frac{\sum_{c\in S}\mu_c^{i,j}}{|S|}. \quad (6)$$

Then $\bar{\mu}_S$ is the vector

$$\bar{\mu}_S = \left[\bar{\mu}_S^{1,2}\bar{\mu}_S^{1,3}\cdots\bar{\mu}_S^{1,K}\bar{\mu}_S^{2,3}\bar{\mu}_S^{2,4}\cdots\bar{\mu}_S^{K-1,K}\right]. \quad (7)$$

In order to more objectively investigate the protein pairs which have higher dependency and are more frequent in cancer samples, a difference of weighted sums was calculated by considering the top N (here set to equal 5 or 10) dependency scores of the ten most frequent phenotypes in each sample. The measure weights the dependency score with the phenotype probability in the sample, and sums all occurrences of the protein pair in all the cancerous samples and of all the normal samples. It then subtracts the score for the normal from the score for the cancer samples, hence giving a positive score if a pair appears more frequently and with higher dependency scores in the cancerous samples. More formally, if $\hat{\mu}_S$ is the vector with the elements of $\bar{\mu}_S$ (lying in [0, 1]) sorted in descending order, p^r_S is the probability of phenotype S in sample r, $S_{\alpha,r}$ is the α^{th} most frequent phenotype in sample r, and

$$M_S^{i,j} = \begin{cases} \bar{\mu}_S^{i,j}, & \text{if } \bar{\mu}_S^{i,j} \text{ is one of the first } N \text{ elements of } \hat{\mu}_S \\ 0, & \text{otherwise} \end{cases}, \quad (8)$$

then the difference of the sum of frequency-weighted localised protein-protein co-dependence/anti-co-dependence values for a protein pair (i, j), $w_{i,j}$ is given by

$$w_{i,j} = \sum_{r\in\psi}\sum_{\alpha=1}^{10} p^r_{S_{\alpha,r}} M^{i,j}_{S_{\alpha,r}} - \sum_{r\in\nu}\sum_{\alpha=1}^{10} p^r_{S_{\alpha,r}} M^{i,j}_{S_{\alpha,r}}. \quad (9)$$

where ψ is the set of cancerous samples, ν is the set of normal samples.

A similar quantity of anti-co-dependence has also been considered by looking at the bottom N dependency scores, so we define

$$\hat{M}_S^{i,j} = \begin{cases} \bar{\mu}_S^{i,j}, & \text{if } \bar{\mu}_S^{i,j} \text{ is one of the last } N \text{ elements of } \hat{\mu}_S \\ 0, & \text{otherwise} \end{cases}, \quad (10)$$

and use $1 - \hat{M}_S^{i,j}$ instead of $M_S^{i,j}$ to measure anti-co-location of protein pairs, i.e.

$$\hat{w}_{i,j} = \sum_{r\in\psi}\sum_{\alpha=1}^{10} p^r_{S_{\alpha,r}}\left(1 - \hat{M}^{i,j}_{S_{\alpha,r}}\right) - \sum_{r\in\nu}\sum_{\alpha=1}^{10} p^r_{S_{\alpha,r}}\left(1 - \hat{M}^{i,j}_{S_{\alpha,r}}\right). \quad (11)$$

Hence, we introduce two new measures called Difference in Sum of Weighted cO-dependence/Anti-co-dependence profiles, further referred to

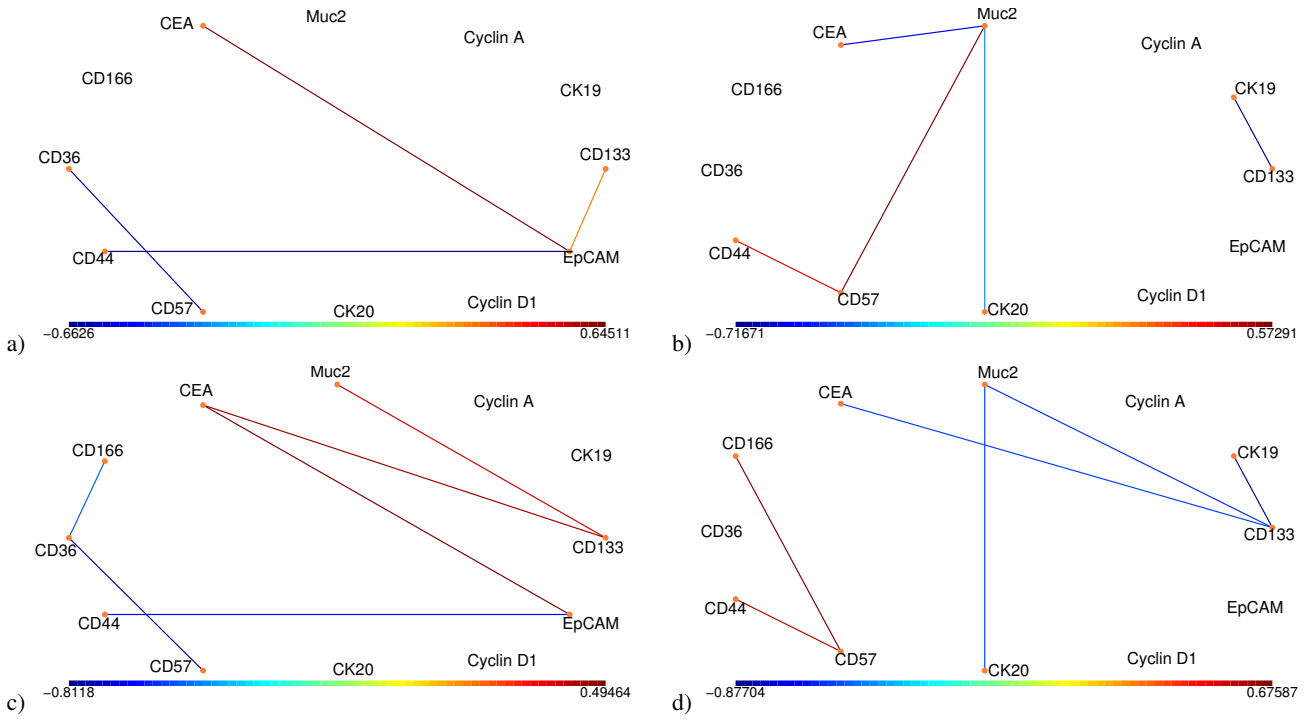

Fig. 3. The social networks of proteins. Each node represents a protein and each edge colour shows a protein pair with different level of co-expression in the normal and cancer samples. Only edges with the top 10% and the bottom 10% of the DiSWOP and DiSWAP values are shown. Figures (a) and (c) show DiSWOP values when considering the top 5 and 10 dependency scores, respectively. Here, a large positive value (shown in red) indicates that the protein pair is more co-dependent in cancer samples, whereas a large negative value (shown in blue) means that the protein pair is more active in normal tissue. Figures (b) and (d) show DiSWAP values when considering the top 5 and 10 dependency scores, respectively. In this case, a large positive value (shown in red) indicates that the protein pair is more anti-co-dependent in cancer samples, whereas a large negative value (shown in blue) means that the protein pair is more anti-co-dependent in normal tissue.

as DiSWOP (Equation 9) and DiSWAP (Equation 11). Large positive values of DiSWOP indicate that the protein pair (i, j) is more co-dependent in cancer samples, while a low negative DiSWOP value means that the protein pair is more co-dependent in the normal samples. Similarly for DiSWAP a large positive value suggests that the protein pair is more anti-co-dependent in cancer and a large negative value that the protein pair is more anti-co-dependent in healthy samples. The DiSWOP and DiSWAP scores are shown in Figure 3. Various combinations of number of phenotypes and dependency scores were also considered. Altering the number of clusters caused very little change to the results as the phenotypes that were added or excluded have very low probability in the samples. On the other hand, increasing the number of dependency scores considerably changed the protein pairs highlighted. However, if more than the top ten scores are included, the average dependency score added to the analysis is below 0.5 and so the proteins are more anti-co-dependent than they are co-dependent. Therefore, these scores should not be included as part of the DiSWOP measure. Further biological validation and analysis of a greater number of samples is needed to determine the optimal number of dependency scores to be considered as part of the dependency measures.

2.5 Synthetic data

The measures presented above were checked using synthetically generated data. Details of the algorithm for generating this data can be found in the Supplementary Materials. Two samples were generated to correspond to one cancer and one normal tissue samples. The expression of 5 tags was simulated for each of these. Each of the samples contained about 80 cells, which were randomly allocated to two different phenotypes per sample, with the first phenotype containing about $1/3$ of the cells in the sample and the rest belonging to the second phenotype. Once the first tag was created, the rest of the tags were generated by keeping a fraction of the pixels the same as in tag 1 and assigning a random value to the rest. The fractions of pixels that were kept the same were as follows:

$$\zeta_c = \begin{bmatrix} 0.4 & 0.6 \\ 0.8 & 0.9 \\ 0.5 & 0.2 \\ 0.1 & 0.6 \end{bmatrix}, \zeta_n = \begin{bmatrix} 0.7 & 0.9 \\ 0.5 & 0.2 \\ 0.6 & 0.4 \\ 0.7 & 0.3 \end{bmatrix}, \tag{12}$$

where ζ_c gives the similarity in the "cancer" sample and ζ_n in the "normal" sample. Each column corresponds to a different phenotype, with the smaller phenotype in the sample being determined by the first column. A row, j in the matrices in Equation 12 gives the similarity between tag 1 and tag $j + 1$. Note that the order of pixels to be kept the same remains constant for a cell, so, for a phenotype S, the prescribed similarity between tags i and j $(i, j > 1)$ is given by $min\left(\zeta(i + 1, S), \zeta(j + 1, S)\right)$. Examples of the images obtained for the "cancer" sample are shown in Figure 5.

3 RESULTS AND DISCUSSION

The results presented in Figure 3 suggest that it is in fact the combinations of protein pairs with high dependency scores that identify cancer cells, which is to be expected, considering the complexity of the system. Calculating the DiSWOP and DiSWAP measures identified pairs which are significantly more co-dependent or anti-codependent in cancer samples than in normal tissue. As

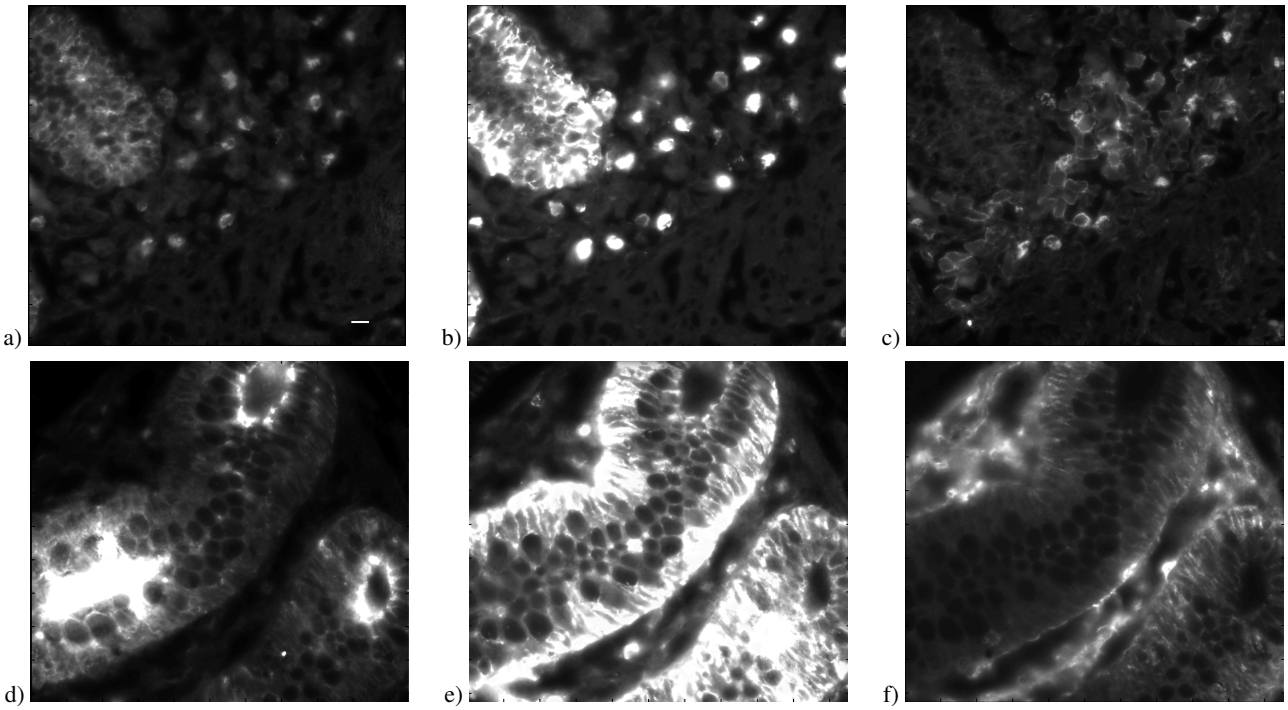

Fig. 4. Protein expression images. Figures (a) - (c) show CEA, EpCAM and CD44 expression levels, respectively, in a cancer sample. Figures (d) - (f) show CEA, EpCAM and CD44 expression levels, respectively, in a normal sample. The scale bar in (a) is 10 μm

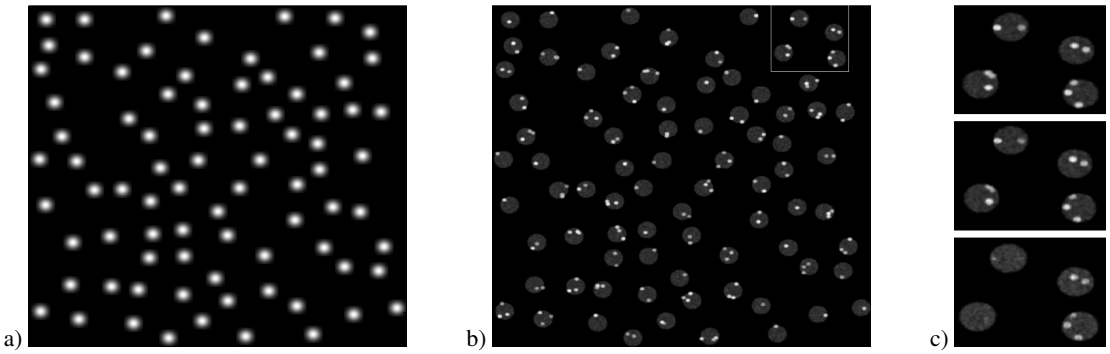

Fig. 5. Example of simulated data. Figures (a) and (b) show the DAPI channel and tag 1, respectively, for the "cancer" sample. Figure (c) shows a zoomed in section, highlighted in Figure (b), of tag 1 (top), tag 3 (middle) and tag 5 (bottom). The two cells on the left hand side belong to the first phenotype and the two cells on the right hand side to the second phenotype.

can be seen in Figure 3 (a) and (c), EpCAM and CEA have very high positive DiSWOP score for both results. This may be due to the fact that both proteins are involved in cell adhesion (details of all the proteins considered have been presented by [Bhattacharya *et al*., 2010]). On the other hand, the pairs CD36 and CD57, and CD44 and EpCAM were more likely to interact in the normal tissue samples (Figures 3 (a) and (c)). These dependencies can be seen in the data. Figure 4 shows the expression levels of CEA, EpCAM and CD44 in a cancer and a normal sample. It is clear that protein expression in Figures 4 (a) and (b) illustrate a higher dependence than in Figures 4 (d) and (e), whereas the expression patterns in Figures 4 (b) and (c) differ more than those in Figures 4 (e) and (f). Similar trends can be seen in most of the other samples. Considering the DiSWAP measure also highlights some pairs of proteins such as CD44 and CD57 being more anti-codependent in cancer samples and Ck19 and CD133 in normal samples. It is important to note that these results were obtained using only 11 samples which, while being a great improvement on previous studies in the toponomics of colon cancer [Bhattacharya *et al*., 2010, Humayun *et al*., 2011], is still insufficient to draw significant biological conclusions. In order to further analyse the consistency of the two dependency measures, the analysis was performed on 16 different combinations of 3 cancer and 3 normal samples. The results are shown in Figure 6 where it can be seen that the protein pairs with highest and lowest DiSWOP and DiSWAP scores are the same as the ones found when all 11 samples were analysed (Figure 3). The protein interactions identified should

be validated biologically once the method has been applied to a large number of samples. However, biological validation could be difficult as the proteins may not interact directly: they are not part of known biological pathways and the observed patterns may be a result of unknown mechanisms in cancer formation.

The use of synthetic data, where the ground truth of the interaction of the tags is known, gives support to the proposed method. The DiSWOP and DiSWAP results for the data generated using Equation 12 are shown in Figure 7. It can be seen that the measures gave the expected results. DiSWOP gave the largest positive value for tags 1 and 3 and the largest negative value for tags 1 and 2. This corresponds to the rows with greatest values in Equation 12. The smaller values of DiSWOP shown in Figure 7 correspond to the second highest values in the matrices in Equation 12. The results for DiSWAP are also as expected – it has identified the tags with lowest similarity in each of the samples.

The framework presented here is novel as it clusters the cells found in a sample, rather than the pixels, as in the methods employed by [Langenkamper *et al*., 2011] and [Humayun *et al*., 2011], and the web-based tool presented by [Kolling *et al*., 2012]. Hence this method enables us to consider the heterogeneity of the samples. Using the MIC scores means that the PPDP is considered rather than the raw expression profile. Therefore, the method is independent of the intensity of the images and hence different stacks can be considered simultaneously. Furthermore, it enables the identification of pairs of proteins which are more active in cancer cells than in normal cells and vice versa. The approach has been developed for images obtained using TIS, but it can also be easily used for other multi-variate imaging techniques, such as MALDI imaging [Cornett *et al*., 2007], Raman microscopy [van Manen *et al*., 2005] and multi-spectral imaging methods [Barash *et al*., 2010].

The proteins used were not chosen because links between them were expected to show up in a protein network, but for a different scientific purpose, namely to help identify cell type. For this reason, relatively few links were considered significant, though with a compensating chance that these links were previously unknown. In the future, we will use additional proteins and we expect to find additional links. Previous work on exploring protein networks in colon cancer have used techniques like microarrays which, unfortunately, destroy all anatomical details. The advantage of our approach is that links in the protein network are found by studying individual cells. A disadvantage, however, is that we are restricted to at most 100 proteins, whereas microarrays measure expression of thousands of genes simultaneously.

The proposed measures could prove more useful once a membrane tag is used to help in a more accurate segmentation of cells. Many of the proteins considered are located in parts of the cell other than the nucleus and these interactions are currently not fully taken into account. Furthermore, a study with an extended tag library may reveal more prominent dependencies specific to cancerous tissue.

[Schubert *et al*., 2006] introduced the ideas of lead and absent proteins in motifs of protein clusters, where a lead protein is one which is present after binarization in all clusters and an absent protein is one which is not present in any of the clusters. These ideas in a way have been expanded by the DiSWOP and DiSWAP measures, which also identify colocation and anti-co-location, respectively. The quantities introduced here provide a measure of the degree, rather than a simple Yes-No classification, of the co-dependence of proteins. Furthermore, they overcome the fact that these proteins are found in both types of tissue by considering the difference between cancer and normal samples.

4 CONCLUSIONS

We have introduced a novel method for analysing multiplex and localised proteomics image data such as the TIS image data. It is different from previously presented methods in that it considers the samples at cell rather than at pixel level, it is intensity independent, and it allows phenotyping of cells based on their protein co-expression profile. Due to the general nature of the framework, the method could be applied to other tissues and/or images obtained from other multivariate imaging techniques. We have presented two new measures of co-dependence and anti-co-dependence, namely DiSWOP and DiSWAP. Applying these over a TIS dataset of eleven samples of cancerous and normal colon tissue, we have found combinations of protein pairs that are much more co-dependent or anti-codependent in cancerous than in normal tissue, pointing to the possibility that combinations of protein pairs rather than single proteins will lead to specific markers for cancer. The results presented here are only preliminary and need to be validated using a larger number of samples and subsequently by other biological techniques. While the number of samples considered is insufficient to draw significant biological conclusions, this is the largest study in colon cancer toponomics conducted to date. Furthermore checks using synthetic data give confidence that our novel measures can help identify and quantify important examples of co-expression and anti-co-expression of protein pairs.

ACKNOWLEDGEMENTS

Funding: V. K.'s research was funded by the BBSRC. A. M. K.'s research was funded by the WPRS. This work is partly funded by the QNRF grant NPRP 5-1345-1-228.

REFERENCES

Adams, J. M. and Strasser, A. (2008) Is tumor growth sustained by rare cancer stem cells or dominant clones, *Cancer Res.*, **68**(11), 4018–4021.

Al-Kofahi, Y. *et al*. (2010) Improved automatic detection and segmentation of cell nuclei in histopathology images. *IEEE Trans Biomed Eng.*, **57**(4), 841–852.

Barash, E. *et al*. (2010) Multiplexed Analysis of Proteins in Tissue Using Multispectral Fluorescence Imaging, *IEEE Transactions on Medical Imaging*, **29** (8), 1457–1462.

Bhattacharya, S. *et al*. (2010) 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 subcellular annotation by using a three symbol code, *J. Proteome Res.*, **9**(12), 6112–6125.

Cornett, D. *et al*. (2007) MALDI imaging mass spectrometry: molecular snapshots of biochemical systems, *Nature Methods*, **4**, 828–33.

Dalerba, P. *et al*. (2007) Phenotypic characterization of human colorectal cancer stem cells. *Proc. Natl. Acad. Sci. USA.* **104**(24), 10158–10163.

Evans, R. G. *et al*. (2012) Toponome imaging system: multiplex biomarkers in oncology. Trends in Molecular Medicine, **18**(12) 723–731.

Frey, B.J. and Dueck, D. (2007) Clustering by Passing Messages Between Data Points, *Science*, **315**, 972–977.

Humayun, A. *et al*., (2011) A Novel Framework for Molecular Co-Expression Pattern Analysis in Multi-Channel Toponome Fluorescence Images, In *MIAAB 2011 (Proceedings Microscopy Image Analysis with Applications in Biology)*, MIAAB, 109–112.

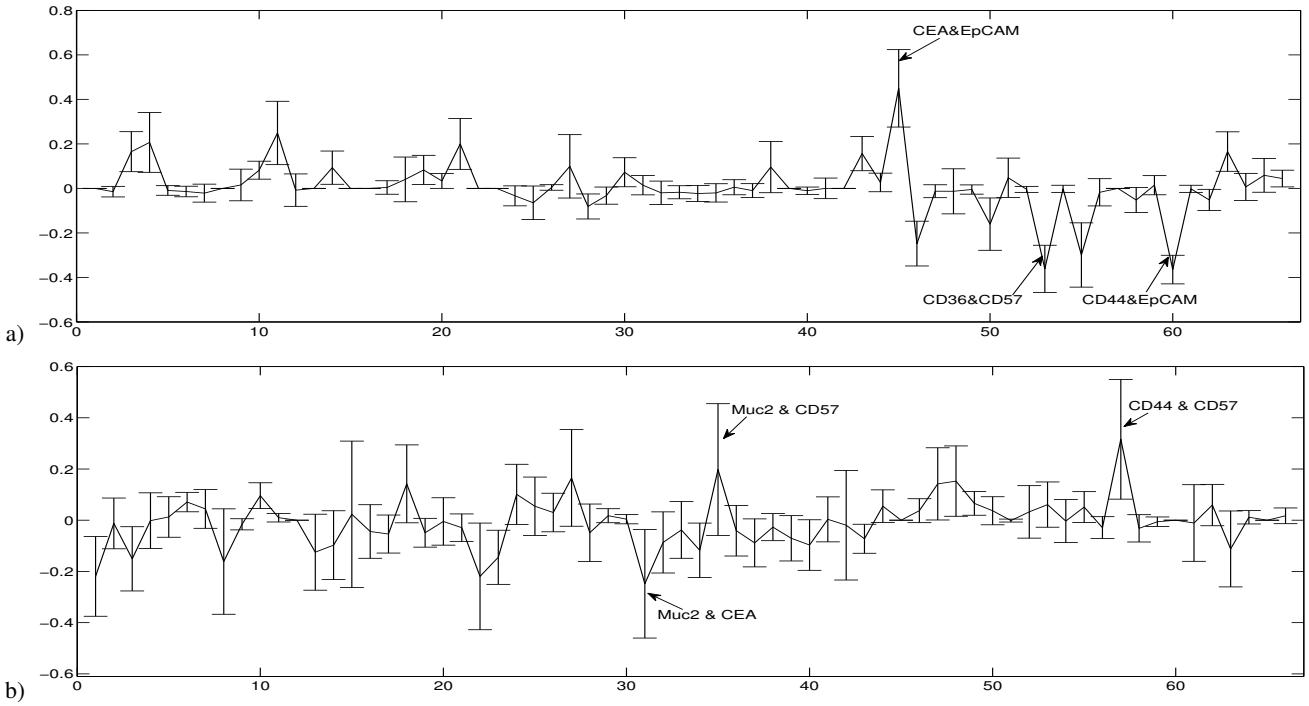

Fig. 6. Mean (a) DiSWOP and (b) DiSWAP values (using the top 5 dependency scores) obtained using 16 different combinations of 3 cancer and 3 normal samples. The error bars are the size of one standard deviation. Numbers along the x-axis correspond to different protein pairs. Note that the labeled protein pairs are the same as the ones highlighted from the analysis of all 11 samples.

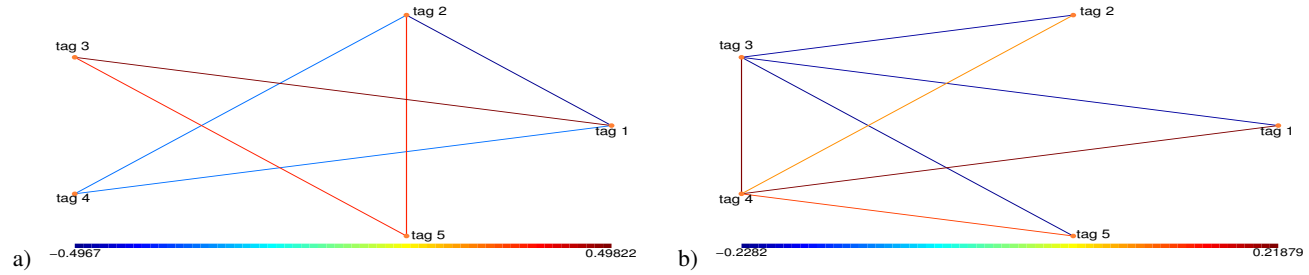

Fig. 7. (a) DiSWOP and (b) DiSWAP values for the simulated data generated using Equation 12.

Khan, A. M. *et al*. (2012) A Novel Paradigm for Mining Cell Phenotypes in Multi-Tag Bioimages using a Locality Preserving Nonlinear Embedding, Lecture Notes in Computer Science, *Neural Information Processing*, Springer Berlin Heidelberg, Vol. 7666, 575–583.

Kolling, J. *et al*., (2012) WHIDE-A web tool for visual data mining colocation patterns in multivariate bioimages, *Bioinformatics*, **28**(8), 1143–1150.

LaBarge, M.A. and Bissell, M. J. (2008) Is CD133 a marker of metastatic colon cancer stem cells, *J. Clin. Invest.* **118**(6), 2021–2024.

Langenkamper, D. *et al*., (2011) Proceedings Workshop on Computational Systems Biology (WCSB) Zurich. Towards protein network analysis using TIS imaging and exploratory data.

Megason, S. and Fraser, S. (2007) Imaging in systems biology, *Cell*, **130**, 784–795.

O'Brien, C.A. *et al*., (2007) A human colon cancer cell capable of initiating tumour growth in immunodeficient mice, *Nature*, **445** (7123), 106–110.

Ontrup, J. and Ritter, H. (2006) Large-scale data exploration with the hierarchically growing hyperbolic SOM, *Neural Networks*, **19**, 751–761.

Pure, E. and Assoian, R.K. (2009) Rheostatic signaling by CD44 and hyaluronan, *Cell Signal*, **21**(5), 651–655.

Raza, S. E. A. *et al*. (2012) RAMTaB: Robust Alignment of Multi-Tag Bioimages, *PLoS ONE*, **7**, e30894.

Reshef, D.N. *et al*. (2011) Detecting Novel Associations in Large Data Sets, *Science*, **334**, 1518–1524.

Ricci-Vitiani, L. *et al*., (2007) Identification and expansion of human colon cancer-initiating cells, *Nature*, **445**, 111–115.

Schubert, W. *et al*., (2003) Topological proteomics, toponomics, MELK-technology. *Adv. Biochem. Eng. Biotechnol.* **83**, 189–209.

Schubert, W. *et al*., (2006) Analyzing proteome topology and function by automated multidimensional fluorescence microscopy, *Nature Biotechnology*, **24**, 1270–1278.

Schubert, W. (2010) On the origin of cell functions encoded in the toponome, *J. Biotechnol.* **149**, 252–259.

Schubert, W. *et al*., (2012) Next-generation biomarkers based on 100-parameter functional super-resolution microscopy TIS. *N. Biotechnol.* **29**, 599–610.

van Manen, H. *et al*. (2005) Single- cell raman and fluorescence microscopy reveal the association of lipid bodies with phagosomes in leukocytes, *PNAS*, **102**(29), 10159–64.

Vucic, E.A. *et al*. (2012) Translating cancer omics to improved outcomes. *Genome Res*. **22**, 188–195.

Weichert, W. *et al*. (2004) ALCAM/CD166 is overexpressed in colorectal carcinoma and correlates with shortened patient survival, *J. Clin. Pathol.*, **57**(11), 1160–1164.