BIOINFORMATICS ORIGINAL PAPER

Vol. 30 no. 3 2014, pages 420–427
doi:10.1093/bioinformatics/btt676

*Bioimage informatics*

Advance Access publication November 21, 2013

# DiSWOP: a novel measure for cell-level protein network analysis in localized proteomics image data

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

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

Associate Editor: Jonathan Wren

## ABSTRACT

**Motivation**: New bioimaging techniques have recently been proposed to visualize 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 analyze the large amounts of multi-tag bioimage data from cancerous and normal tissue specimens to begin to infer protein networks and unravel the cellular heterogeneity at a molecular level.

**Results**: The proposed approach analyzes cell phenotypes in normal and cancerous colon tissue imaged using the robotically controlled Toponome Imaging System microscope. It involves segmenting the 4',6-diamidino-2-phenylindole-labelled image into cells and determining the cell phenotypes according to their protein–protein dependence profile. These were analyzed 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 Toponome Imaging System image stacks from either cancerous or normal human colorectal specimens. This approach enables one to easily identify protein pairs that have significantly higher/lower co-expression levels in cancerous tissue samples when compared with normal colon tissue.

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

**Contact:** v.n.kovacheva@warwick.ac.uk or Nasir.Rajpoot@ieee.org

**Supplementary Information:** Supplementary data are available at *Bioinformatics* online.

Received on July 4, 2013; revised on September 30, 2013; accepted on November 17, 2013

## 1 INTRODUCTION

To understand cellular biology on a systems level, relationships between molecular components must be understood not only at a functional level but also localized 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 visualize the colocation or interaction of several proteins in cells in intact tissue specimen. These include Matrix-assisted laser desorption/ionization (MALDI) imaging (Cornett *et al.*, 2007), Raman microscopy (Van Manen *et al.*, 2005), Toponome Imaging System (TIS) (Schubert *et al.*, 2006), multi-spectral imaging methods (Barash *et al.*, 2010) and MxIF (Gerdes *et al.*, 2013). TIS is an automated high-throughput technique able to co-map up to a 100 different proteins or other tag-recognizable biomolecules 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, co-dependence between two proteins is a potential indication for an interaction that is not necessarily direct. TIS has a sub-cellular maximum lateral resolution of 206 × 206 nm/pixel (Bhattacharya *et al.*, 2010), which allows the determination of sub-cellular protein network architectures. The combination of proteomic information with spatial sub-cellular level topographical data has been termed 'toponomics' (Schubert *et al.*, 2003, 2012).

Biomarkers used in current clinical practice are limited to the simultaneous analysis of only a handful of proteins. Therefore, they 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 personalized medicine. Studying the protein interactions in cancer could uncover previously unknown mechanisms of tumour formation and could identify new potential drug targets in the form of protein interactions.

It has been shown that TIS imaging can be used in cancer research for protein network mapping (Bhattacharya *et al.*, 2010). However, there is a need for new algorithmic approaches that analyze the co-expression patterns. The standard way to analyze TIS images is to apply a threshold to each image of the stack and thus reduce it to binary values (Schubert *et al.*, 2006). However, while this step is straightforward and can be performed objectively (Barysenka *et al.*, 2010), by reducing the image to binary, a lot of potentially important information is

*To whom correspondence should be addressed lost. Recently, such non-threshold methods have been presented (Humayun *et al.*, 2011; Langenkamper *et al.*, 2011). These algorithms cluster molecular co-expression patterns 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). Although this tool is effective at identifying molecular co-expression patterns, the cellular structure is lost, and hence the method is unable to analyze the different cell phenotypes that may be present in the samples. More recently, focus has shifted towards cell-level analysis. One approach is to use K-median clustering to phenotype-segmented cells (Gerdes *et al.*, 2013). However, the conclusions in this study about pathways in colon cancer were drawn by only visual inspection of the phenotypes obtained and without considering any control samples. In another study (Khan *et al.*, 2013), cells were phenotyped after dimensionality reduction of their raw expression vector using *t*-distributed stochastic neighbor embedding (Van der Maaten and Hinton, 2008).

In this article, 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 downregulation of proteins, is (or can be) the 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. Finally, two new measures are proposed to enable us to infer small-scale protein networks. These new measures highlight protein pairs that have different interaction in cancer and normal tissues and, hence, can be used as biomarkers to distinguish between different types of samples using protein co-dependence. To our knowledge, this is the first study of localized protein networks performed using multi-labelling imaging techniques. 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.

## 2 METHODS

### 2.1 Data and pre-processing

The image data used in this study were 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 10 cm away from the visible margin of the tumour. Two visual fields were manually selected in each tissue sample, resulting in four TIS datasets from a single patient. The results presented here were obtained by considering 11 samples—6 healthy and 5 cancerous. A library of 12 antibody tags (see Supplementary Material for a list of criteria on how these were selected) was used based on recent findings (Bhattacharya *et al.*, 2010). Some of the tags are known tumour markers or cancer stem cell markers. These were CD133, CK19, Cyclin A, Muc2, CEA, CD166, CD36, CD44, CD57, CK20, Cyclin D1 and EpCAM. The stacks also included a 4',6-diamidino-2-phenylindole (DAPI) tag used to identify the cell nuclei. A previously presented protocol for sample preparation and image acquisition was used (Bhattacharya *et al.*, 2010).

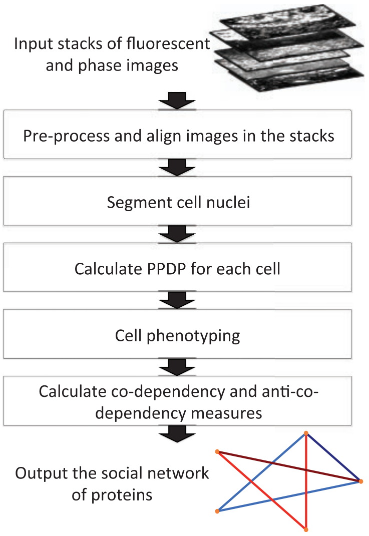
Input stacks of fluorescent and phase images
Pre-process and align images in the stacks
Segment cell nuclei
Calculate PPDP for each cell
Cell phenotyping
Calculate co-dependency and anti-co-dependency measures
Output the social network of proteins

**Fig. 1.** Overview of the proposed framework

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 to prevent possible noise resulting from the slight misalignment of the multi-tag images obtained using TIS. This method has been shown to achieve sub-pixel accuracy of registering these data (Raza *et al.*, 2012). Then, if there are $K$ tags, each having a corresponding image of size $m \times n$, the data can be represented as a $K \times mn$ matrix:

$$\mathbf{X} = \begin{bmatrix} x^1_{1,1} & x^1_{1,2} & \cdots & x^1_{m,n} \\ x^2_{1,1} & x^2_{1,2} & \cdots & x^2_{m,n} \\ \vdots & \vdots & \ddots & \vdots \\ x^K_{1,1} & x^K_{1,2} & \cdots & x^K_{m,n} \end{bmatrix} \tag{1}$$

where $x^k_{i,j}$ is the expression level of protein $k$ ($k < K$) at pixel $(i, j)$. In our experiments, $K = 12, m = 1027$ and $n = 1056$.

We recently proposed to perform cell segmentation of TIS stacks to restrict the analysis to cellular areas only (Khan *et al.*, 2012). 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, as 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 used in this set of experiments to enable 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) applied to a DAPI channel (Khan *et al.*, 2012). This was necessary 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 and can be used as a rough approximation of the cells. Details of the method and examples can be found in the Supplementary Materials.

Segmentation resulted in 2945 cells being identified. The cell-localized protein expression values for each of the $K$ proteins are collected in a protein expression matrix $\mathbf{X}_c$ of the order $K \times N_c$ for each cell $c$

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

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 $c$ 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:

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

## 2.2 Protein–protein dependence profile

The pairwise maximal information coefficient (MIC) (Reshef *et al.*, 2011) for each pair of proteins, localized to an individual cell $c$, is calculated to obtain the PPDP of the cell. We used this statistic, as 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 $\mu_c$ of pairwise MIC scores is obtained. The vector represents the PPDP of the cell and can be expressed as

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

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 Figure 2.

Other co-dependence measures were also considered for the analysis. Pearson and Spearman correlations fail to capture non-linear relationships between protein expression profiles, which often occur because of the inhomogeneous structure of the cells. Mutual information and normalized mean expression values were also tested. However, each of these resulted in a batching effect where some phenotypes were predominantly located in a single, usually cancerous, sample (see Section 3 of Supplementary Materials). 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 that functionality can be determined by colocation rather than changes in abundance levels (Schubert, 2010). The analysis was also performed using distance correlation (Szkely and Rizzo, 2009). The final results obtained were similar to the ones obtained using MIC (see Supplementary Materials). However, it has been found that the distance correlation has a strong preference for some types of dependencies and gives different scores at the same noise levels (Reshef *et al.*, 2011). Therefore, the MIC is preferred due to its robustness to variations in the type of dependence.

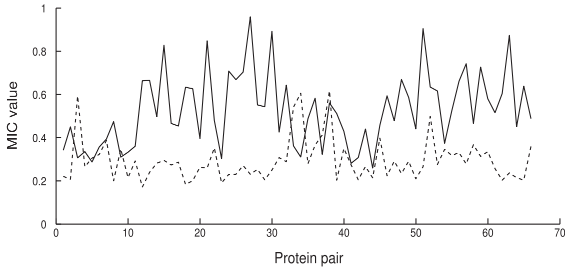

**Fig. 2.** PPDP of two cells from the same specimen

## 2.3 Cell phenotyping based on localized PPDP

The vector $\mu_c$ is the PPDP of the cell $c$ and can be used to determine the cell phenotype using a clustering algorithm. Affinity propagation 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 $\mu_a$ and $\mu_b$, respectively, the $(a,b)$ entry of the similarity matrix (for $a \neq b$) is given by

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

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. An agglomerative hierarchical clustering (Jain *et al.*, 1999) and Gaussian Bayesian hierarchical clustering (Kovacheva *et al.*, 2014; Sirinukunwattana *et al.*, 2013) approaches with the same number of clusters were also considered. It was encouraging to see that these gave similar results, which are shown in the Supplementary Materials.

## 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 $i<j \le K$) $\bar{\mu}_S^{i,j}$ is given by

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

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

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

To more objectively investigate the protein pairs that 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 10 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. The sums are normalized by the number of 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, $p_S^r$ is the probability of phenotype $S$ in sample $r$, $S_{\alpha,r}$ is the $\alpha^{th}$ most frequent phenotype in sample $r$, and

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

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

$$w_{i,j} = \frac{1}{|\psi|}\sum_{r\in\psi}\sum_{\alpha=1}^{10} p^{r}_{S_{\alpha,r}} M^{i,j}_{S_{\alpha,r}} - \frac{1}{|\nu|}\sum_{r\in\nu}\sum_{\alpha=1}^{10} p^{r}_{S_{\alpha,r}} M^{i,j}_{S_{\alpha,r}} \tag{9}$$

where $\psi$ is the set of cancerous samples and $\nu$ 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

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

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

$$\begin{aligned} \hat{w}_{i,j} = & \frac{1}{|\psi|}\sum_{r\in\psi}\sum_{\alpha=1}^{10} p^{r}_{S_{\alpha,r}}\left(1 - \hat{M}^{i,j}_{S_{\alpha,r}}\right) \\ & - \frac{1}{|\nu|}\sum_{r\in\nu}\sum_{\alpha=1}^{10} p^{r}_{S_{\alpha,r}}\left(1 - \hat{M}^{i,j}_{S_{\alpha,r}}\right) \end{aligned} \tag{11}$$

Hence, we introduce two new measures called Difference in Sum of Weighted cO-dependence/Anti-co-dependence profiles, further referred to 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, whereas 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 suggests 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 little change to the results, as the phenotypes that were added or excluded have 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 10 scores are included, the average dependency score added to the analysis is <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 these data can be found in the Supplementary Materials. Initially, two samples were generated to correspond to one cancer and one normal tissue samples. The expression of five tags was simulated for each of these samples. Each of the samples contained ~80 cells, which were randomly allocated to two different phenotypes per sample, with the first phenotype containing about one-third 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 $\zeta_c$ gives the similarity in the 'cancer' sample and $\zeta_n$ 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 $min(\zeta(i+1, S), \zeta(j+1, S))$. Examples of the images obtained for the 'cancer' sample are shown in Figure 4.

A larger simulation was run to test the power of the DiSWOP measure for discriminating between cancer and normal samples. This consisted of generating 30 samples with six tags and five initialized phenotypes for each sample type (see Supplementary Materials for details). The DiSWOP values were calculated on a training set of five cancer and five normal samples. The protein pair that gave the highest DiSWOP value was selected as a biomarker. Then, each sample in the test set was individually phenotyped and the DiSWOP values were calculated, treating the sample as cancerous in order to obtain non-negative scores. In the cancerous samples, the DiSWOP values for the selected protein pair had a mean of 0.28 and a standard deviation of 0.06, whereas in the normal samples, the values had a mean of 0.04 and a standard deviation of 0.04. Therefore, in this set of synthetic data, the DiSWOP measure provided an easy way to discriminate between the types of samples. In addition, it was encouraging to see that the phenotyping results were similar to the initialized phenotypes. Although most initialized phenotypes were subdivided into two or three phenotypes, this was expected with the synthetic data due to the large number of random pixel intensities introduced when generating the expression of the tags. Only ~2% of the cells were misclassified by the affinity propagation clustering.

# 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-co-dependent in cancer samples than in normal tissue. As can be seen in Figure 3a and c, EpCAM and CEA have 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 previously presented (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 (Fig. 3a and c). These dependencies can be seen in the data. Figure 5 shows the expression levels of CEA, EpCAM and CD44 in a cancer and a normal sample. It is clear that protein expression in Figure 5a and b illustrate a higher dependence than in Figure 5d and e, whereas the expression patterns in Figure 5b and c differ more than those in Figure 5e 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-co-dependent 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. To further analyze the consistency of the two dependency measures, the analysis was performed on 16 different combinations of three cancer and three 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

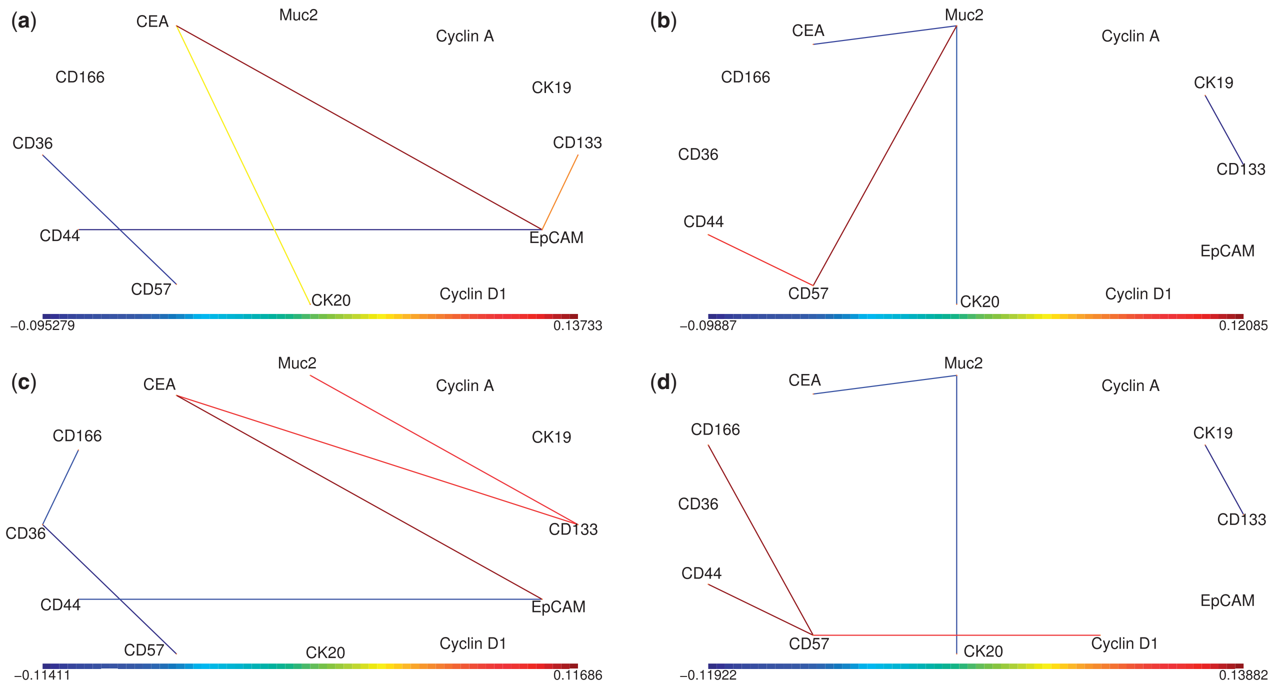

**Fig. 3.** The social networks of proteins. Each node represents a protein and each edge colour shows a protein pair with different level of co-expression in the normal and cancer samples. Only edges with the top 10% and the bottom 10% of the DiSWOP and DiSWAP values are shown. (**a**) and (**c**) show 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. (**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

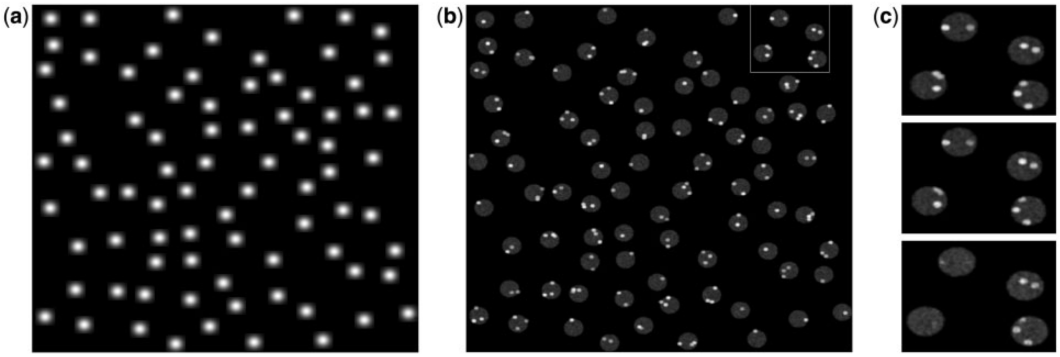

**Fig. 4.** Example of simulated data. (**a**) and (**b**) show the DAPI channel and tag 1, respectively, for the ‘cancer’ sample. (**c**) shows a zoomed in section, highlighted in (b), of tag 1 (top), tag 3 (middle) and tag 5 (bottom). The two cells on the left side belong to the first phenotype, and the two cells on the right side to the second phenotype

samples were analyzed (Fig. 3). The protein interactions identified should be validated biologically once the method has been applied to a large number of samples. In fact, there are pathways involving some of the protein pairs highlighted, which have been found to play an important role in colorectal cancer (Kovacheva *et al.*, 2014).

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.

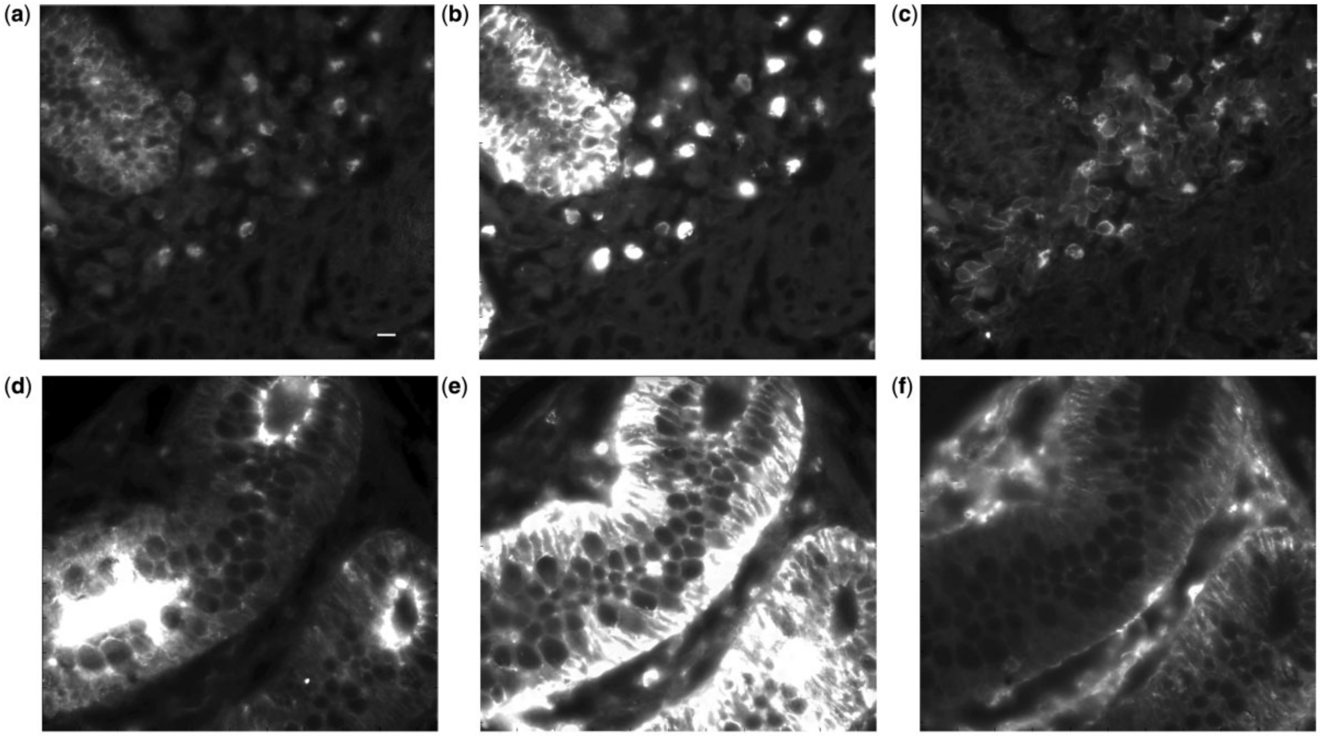

**Fig. 5.** Protein expression images. (**a–c**) show CEA, EpCAM and CD44 expression levels, respectively, in a cancer sample. (**d–f**) show CEA, EpCAM and CD44 expression levels, respectively, in a normal sample. The scale bar in (a) is 10 $\mu$m

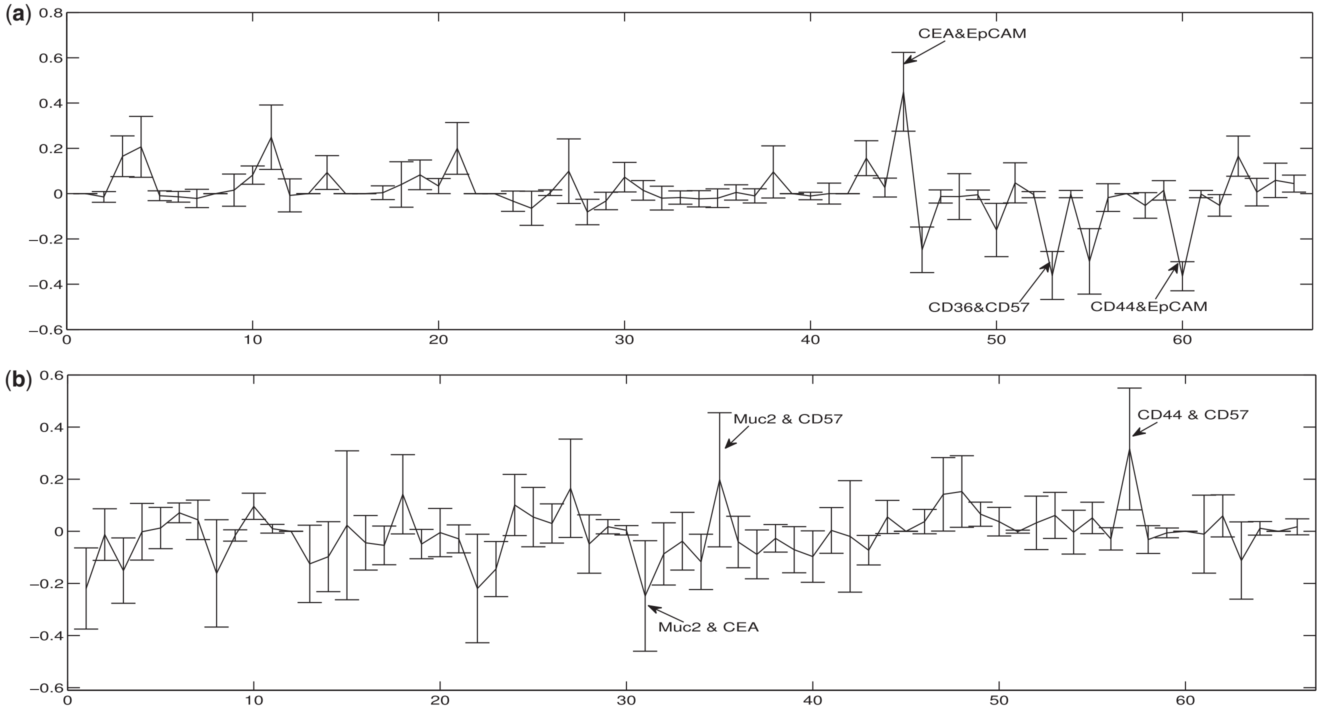

**Fig. 6.** Mean (**a**) DiSWOP and (**b**) DiSWAP values (using the top 5 dependency scores) obtained using 16 different combinations of 3 cancer and 3 normal samples. The error bars are the size of one standard deviation. Numbers along the *x*-axis correspond to different protein pairs. Note that the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

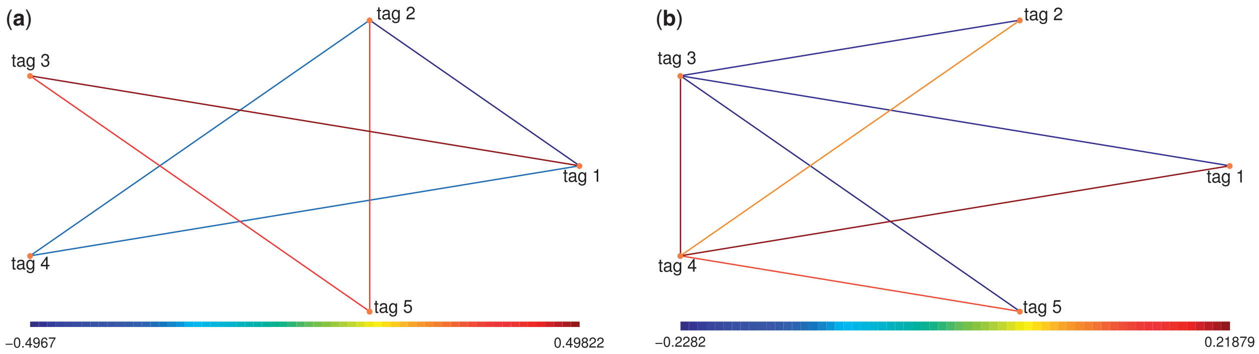

**Fig. 7.** (**a**) DiSWOP and (**b**) DiSWAP values for the simulated data generated using Equation (12)

The framework presented here is novel, as it clusters the cells found in a sample, rather than the pixels, as in previous methods (Humayun *et al.*, 2011, Kolling *et al.*, 2012; Langenkamper *et al.*, 2011). 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 that 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 multivariate imaging techniques, such as MALDI imaging (Cornett *et al.*, 2007), Raman microscopy (Van Manen *et al.*, 2005) and multispectral 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, although 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 works 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 the links in the protein network are found by studying individual cells. However, a disadvantage 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.

The binarization method (Schubert *et al.*, 2006) introduced the ideas of lead and absent proteins in motifs of protein clusters, where a lead protein is one that is present after binarization in all clusters and an absent protein is one that 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 analyzing multi-label 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. Owing 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 11 samples of cancerous and normal colon tissue, we have found combinations of protein pairs that are much more co-dependent or anti-co-dependent 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. Although the number of samples considered is insufficient to draw significant biological conclusions, this is the largest study of colon cancer using TIS conducted to date. Furthermore, checks using synthetic data give confidence that our novel measures can help identify and quantify important examples of co-dependence and anti-co-dependence of protein pairs.

*Funding*: V.K. is fully funded by the UK-based Biotechnology and Biological Sciences Research Council (BBSRC). A.M.K. is partly funded by the Warwick Postgraduate Research Scholarship (WPRS) and partly by the Department of Computer Science at Warwick. This publication was made possible by NPRP grant # NPRP 5-1345-1-228 from the Qatar National Research Fund (a member of the Qatar Foundation). The statements made herein are solely the responsibility of the authors.

*Conflict of Interest*: none declared.

## REFERENCES

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

Barash,E. *et al.* (2010) Multiplexed analysis of proteins in tissue using multispectral fluorescence imaging. *IEEE Trans. Med. Imaging*, **29**, 1457–1462.

Barysenka,A. *et al.* (2010) An information theoretic thresholding method for detecting protein colocalizations in stacks of fluorescence images. *J. Biotechnol.*, **149**, 127–131.

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**, 6112–6125.

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

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

Frey,B.J. and Dueck,D. (2007) Clustering by passing messages between data points. *Science*, **315**, 972–977.

Gerdes,M.J. *et al.* (2013) Highly multiplexed single-cell analysis of formalin- fixed, paraffin-embedded cancer tissue. *Proc. Natl Acad. Sci. USA*, **110**, 11982–11987.

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)*. pp. 109–112.

Jain,A. *et al.* (1999) Data clustering: a review. *ACM Comput. Surv.*, **31**, 264–323.

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. In: *Neural Information Processing*. Vol. 7666, Springer, Berlin Heidelberg, pp. 575–583.

Khan,A.M. *et al.* (2013) Cell phenotyping in multi-tag fluorescent bioimages. *Neurocomputing*.

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

Kovacheva,V. *et al.* (2014) A bayesian framework for cell-level protein network analysis for multivariate proteomics image data. In: *SPIE Med. Imaging*. San Diego, USA.

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

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

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

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.

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

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. *Nat. Biotechnol.*, **24**, 1270–1278.

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

Sirinukunwattana,K. *et al.* (2013) Bayesian hierarchical clustering for studying cancer gene expression data with unknown statistics. *PLoS One*, **8**, e75748.

Szkely,G.J. and Rizzo,M.L. (2009) Brownian distance covariance. *Ann. Appl. Stat.*, **3**, 1236–1265.

Van der Maaten,L. and Hinton,G. (2008) Visualizing data using t-sne. *J. Mach. Learn. Res.*, **9**, 2579–2605.

van Manen,H. *et al.* (2005) Single- cell raman and fluorescence microscopy reveal the association of lipid bodies with phagosomes in leukocytes. *Proc. Natl Acad. Sci. USA*, **102**, 10159–10164.

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