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

Kovacheva et al.

1 Data

The data used was obtained from 26 cycles of the TIS machine. However, some of these were excluded from the analysis using the following criteria:

1. Function of the tag not relevant to the study - this way we excluded 2 DAPI channels with different tag concentrations and 5 PBS runs, which were performed to remove autofluorescence.

2. Tag was not registered properly by the RAMTaB algorithm [1]. A measure of confidence in the registration results is given by the standard deviation of shifts computed by different blocks and images were discarded if this exceeded a pre-defined threshold as described in [1].

3. Invalid expression - images were checked by a pathologist to validate expression of the protein tag in the image. This resulted in all images of the Ki67 tag to be excluded. This protein is expected to be found in increased concentrations only in proliferating cells but this was not the case (Figure 1).

2 Segmentation

Each image was segmented using a modified form of the graph cut method [2] proposed by our group [3] applied to a DAPI channel (Figure 2). Initially, each image is binarised using graph-cut based algorithm to extract the foreground. Next, an initial segmentation is performed by detecting seed points on the foreground of the binarised image by using a multi-scale Laplacian of Gaussian (LoG) filter [4]. The initial segmentation is then refined using a second graph-cut based algorithm. Finally, the nuclei segmentation results obtained using the framework are post-processed by either eliminating very small nuclei or merging them with nearby nuclei, as they usually result from segmentation errors. This final step ensures that analysis is restricted only to clearly distinguishable nuclei. This serves as a rough approximation of the pixels belonging to the cells. Segmentation is currently still an issue as gold-standard data is not available and perfect cutting of sections is impossible. However, future experiments will include a membrane tag, which should resolve this problem.

3 Protein-Protein Dependence Profile (PPDP)

The pairwise maximal information coefficient (MIC) [5] for each pair of proteins, localised to an individual cell c, is calculated to obtain the protein-protein dependence profile (PPDP) of the cell. In order to do this the intensities of the two proteins are considered pixel by pixel. The MIC is calculated by exploring all grids on the scatterplot up to a maximal grid resolution dependent on the grid size, computing for every pair of integers \((k,l)\) the largest possible mutual information achievable by any \(k\)-by-\(l\) grid applied to the data. The values found are then normalised as follows: for a grid \(G\), let \(I_G\) denote the mutual information of the probability distribution induced on the grid boxes of \(G\), where the probability of a box is proportional to the
number of points that fall in the box. The highest normalised mutual information achieved by any $k$-by-$l$ grid is recorded as the element $m_{k,l}$ of a characteristic matrix $M$, where

$$m_{k,l} = \frac{\max(G)}{\log(\min\{k,l\})},$$

(1)

with the maximum being taken over all $k$-by-$l$ grids $G$. The normalisation ensures a fair comparison between grids of different sizes and obtains values between 0 and 1. The MIC is the maximum value of $M$ [5]. As suggested by [5], the maximum size of the grids considered was set to be $kl < N_c^{0.6}$ where $N_c$ is the number of pixels in the cell $c$.

Other dependency measures have been considered. Linear measures, such as Pearson’s and Spearman’s coefficient, have been found unsuitable as some cells show non-linear dependence. An example of this can be seen in Figure 3. In Figure 3 (a) we can see that the two proteins are weakly dependent on each other. However, the Pearson’s coefficient for this cell was -0.01, whereas the MIC was 0.33. Mutual information and normalised 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 and the samples were split into a handful of phenotypes (Figure 4). For result comparison distance correlation (DC) [6] was also used. While this measure gives comparable results (See Table 1), it has been found that the DC has a strong preference for some types of dependencies and gives different scores at the same noise levels [5]. Therefore, the MIC is preferred due to its robustness to variations in the type of dependence.

4 Generating Synthetic Data

Synthetic data was generated in order to test our methods. The advantage with synthetic data is that the ground truth is known, independently of the analytical method.

To generate data with $K$ tags and $P$ phenotypes we follow the following algorithm:

Input: a similarity matrix, $\zeta$, which is a $(K - 1) \times P$ matrix, whose $j$-th row contains, for each phenotype, a similarity value (in $[0, 1]$) between tag 1 and tag $j + 1$; a phenotype ratio, $\phi$, which is an integer vector of length $P$, specifying the proportion of cells to be contained within each phenotype (e.g. if $P = 2$ and $\phi = [1, 2]$, 1/3 of the cells will have the first phenotype); number of cells desired, $N$.

1: procedure Synthetic Data($\zeta, \phi, N$)
Figure 2: Segmentation results on a part of a normal sample (a) and a part of a cancer sample (b). The size of the scale bars is 10 µm.
Table 1: Top and bottom 10 DiSWOP results from different dependency measures (MIC and DC) and clustering methods (AP, Gaussian Bayesian Hierarchical clustering (GBHC) and Agglomerative Hierarchical Clustering (AHC)). Pairs are shown with decreasing DiSWOP score. All results have been obtained by considering the top 5 PPDP scores for each phenotype.

<table>
<thead>
<tr>
<th>MIC and AP</th>
<th>MIC and GBHC</th>
<th>MIC and AHC</th>
<th>DC and AP</th>
</tr>
</thead>
<tbody>
<tr>
<td>CEA &amp; EpCAM</td>
<td>CK20 &amp; EpCAM</td>
<td>CEA &amp; EpCAM</td>
<td>CEA &amp; EpCAM</td>
</tr>
<tr>
<td>CD133 &amp; EpCAM</td>
<td>CEA &amp; EpCAM</td>
<td>CD133 &amp; CK20</td>
<td>CD133 &amp; Mic2</td>
</tr>
<tr>
<td>CEA &amp; CK20</td>
<td>Mic2 &amp; EpCAM</td>
<td>CK19 &amp; CK20</td>
<td>CK19 &amp; CEA</td>
</tr>
<tr>
<td>CD133 &amp; Muc2</td>
<td>CD133 &amp; CEA</td>
<td>CK19 &amp; EpCAM</td>
<td>CK19 &amp; EpCAM</td>
</tr>
<tr>
<td>CD133 &amp; CEA</td>
<td>Mic2 &amp; CEA</td>
<td>CK19 &amp; CEA</td>
<td>CK19 &amp; CEA</td>
</tr>
<tr>
<td>CK19 &amp; EpCAM</td>
<td>CK19 &amp; CEA</td>
<td>Mic2 &amp; EpCAM</td>
<td>CD133 &amp; EpCAM</td>
</tr>
<tr>
<td>CK20 &amp; EpCAM</td>
<td>CD133 &amp; Muc2</td>
<td>CD57 &amp; EpCAM</td>
<td>Cyclin A &amp; CD57</td>
</tr>
<tr>
<td>CD133 &amp; Cyclin D1</td>
<td>CK19 &amp; CK20</td>
<td>CEA &amp; CK20</td>
<td>CD133 &amp; CEA</td>
</tr>
<tr>
<td>Muc2 &amp; EpCAM</td>
<td>CEA &amp; CK20</td>
<td>CD133 &amp; CEA</td>
<td>Mic2 &amp; EpCAM</td>
</tr>
<tr>
<td>CD57 &amp; EpCAM</td>
<td>CK19 &amp; EpCAM</td>
<td>Cyclin A &amp; EpCAM</td>
<td>CD57 &amp; EpCAM</td>
</tr>
<tr>
<td>CD166 &amp; Cyclin D1</td>
<td>Cyclin A &amp; CD57</td>
<td>CD44 &amp; CK20</td>
<td>CD57 &amp; Cyclin D1</td>
</tr>
<tr>
<td>Cyclin A &amp; CK20</td>
<td>CD133 &amp; Cyclin D1</td>
<td>Muc2 &amp; CD44</td>
<td>Mic2 &amp; CD44</td>
</tr>
<tr>
<td>CD166 &amp; CD57</td>
<td>CD166 &amp; CD57</td>
<td>Muc2 &amp; CD166</td>
<td>Mic2 &amp; CD36</td>
</tr>
<tr>
<td>Muc2 &amp; CD44</td>
<td>CD57 &amp; Cyclin D1</td>
<td>CD57 &amp; Cyclin D1</td>
<td>Mic2 &amp; CD57</td>
</tr>
<tr>
<td>Muc2 &amp; CD166</td>
<td>CD166 &amp; CD36</td>
<td>CK19 &amp; CD57</td>
<td>Cyclin A &amp; CK20</td>
</tr>
<tr>
<td>CK19 &amp; Cyclin A</td>
<td>Cyclin A &amp; CD166</td>
<td>CD166 &amp; CD36</td>
<td>CD166 &amp; CD36</td>
</tr>
<tr>
<td>CD166 &amp; CD36</td>
<td>CD44 &amp; EpCAM</td>
<td>CD166 &amp; Cyclin D1</td>
<td>CD166 &amp; EpCAM</td>
</tr>
<tr>
<td>CD36 &amp; Cyclin D1</td>
<td>CD166 &amp; Cyclin D1</td>
<td>CD36 &amp; CD57</td>
<td>CD36 &amp; EpCAM</td>
</tr>
<tr>
<td>CD36 &amp; CD57</td>
<td>CD36 &amp; Cyclin D1</td>
<td>CD36 &amp; Cyclin D1</td>
<td>CD36 &amp; Cyclin D1</td>
</tr>
<tr>
<td>CD44 &amp; EpCAM</td>
<td>CD36 &amp; CD57</td>
<td>CD44 &amp; EpCAM</td>
<td>CD36 &amp; CD57</td>
</tr>
</tbody>
</table>
Figure 3: An example of non-linear dependence between protein expressions in a cell. Figure (a) shows a scatter plot of the pixel intensities of CK19 and CD166 in a cancer cell. Figures (b) - (d) show the DAPI, CK19 and CD166 expression of the cell outlined in red.

2: \( \text{coor} \leftarrow 2N \times 2 \text{ matrix with random coordinates for the centres of cell nuclei} \)
3: \( \text{Remove cells that are overlapping and update the value of } N \)
4: \( \text{DAPI(coor}(\cdot, 1), \text{coor}(\cdot, 2)) \leftarrow 1 \) (0 elsewhere)
5: \( \text{DAPI} \leftarrow \text{convolve image with 2D Gaussian (}\sigma = \text{cellRadius}/2\) )

▷ Plot DAPI image (Figure 5 (a))
6: \( \text{type} \leftarrow P \times N \text{ matrix with randomly assigned cell phenotype indication satisfying the ratios in } \phi \)
7: \( \text{tags} \leftarrow \text{empty height} \times \text{width} \times K \text{ matrix} \) ▷ Expression of tags
8: \( N_p \leftarrow \text{Number of pixels in a cell} \)
9: \( \zeta_p \leftarrow \lfloor \zeta * N_p \rfloor \) ▷ Number of pixels that remain the same
10: for cell \( \leftarrow 1 : N \) do

11: Pick \( N_s \) from the set \{1, 2, 3\} ▷ Number of bright spots in the cell
12: for Spot \( \leftarrow 1 : N_s \) do ▷ Create bright spots
13: \( [xc, yc] \leftarrow \text{Generate random coordinates for the spot center} \)
14: Add a disk, centred at \( \text{tags}(xc, yc, 1) \) with radius 5 pixels and a random brightness in \( [0.5, 1] \)
15: end for
16: Add a random value in \( [0, 0.5] \) to each pixel of the cell in tag 1
17: \( \gamma \leftarrow N_p \times \text{matrix containing a random permutation of the coordinates of the pixels within the cell} \)
18: \( \phi \leftarrow \text{Cell phenotype from type} \)
19: for tag \( \leftarrow 2 : K \) do ▷ Generate other tags
20: \( \gamma_p \leftarrow \text{First } \zeta_p (\text{tag} - 1, \phi) \text{ elements of } \gamma \)
21: for \( [xp, yp] \leftarrow \text{a pixel coordinates in the cell} \) do
22: if \( [xp, yp] \in \gamma_p \) then
23: \( \text{tags}(xp, yp, \text{tag}) \leftarrow \text{tags}(xp, yp, 1) \)
24: else \( \text{tags}(xp, yp, \text{tag}) \leftarrow \text{new random value from } [0, 0.5] \)
Figure 4: Distribution of phenotypes obtained using affinity propagation (AP) based on the mutual information profile of the cells amongst (a) cancerous samples and (b) normal samples. Each colour corresponds to a different sample. The 41 phenotypes are shown along the $x$-axis. The $y$-axis shows proportion of the phenotype located in each sample.
Figure 5: Example of simulated data generated by Equation (12). Figure (a) shows the DAPI channel and Figures (b) - (f) show tags 1 to 5, respectively, for the "cancer" sample.

```plaintext
25:     end if
26:     end for
27:    end for
28:  end for
29:  Normalise each tag to [0, 1] and convolve with 2D Gaussian ($\sigma = 2$)
    ▶ Represents the point spread function (Figure 5 (b) - (f))
30: end procedure

In the extended simulation, all 30 samples were generated with the following similarity matrices:
```
Figure 6: Top and bottom 3 DiSWOP results from training data in the extended simulation.

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

(2)

where \( \zeta_c \) gives the similarity in the “cancer” samples and \( \zeta_n \) in the “normal” samples. Each column corresponds to a different phenotype with a phenotype ratio given by \( \phi = [2, 1, 1, 1, 1] \). The top and bottom 3 DiSWOP results for the training data can be seen in Figure 6.

References