Cancer Image Analysis

Advances in Discovery of Complex Biomarkers for Colorectal Cancer Using Multiplexed Proteomics Imaging

Abstract
Multiplexed proteomics imaging techniques such as the Toponome Imaging System (TIS) can yield high-resolution images of multiple proteins co-localised within individual cells. This enables the study of protein interactions and tumour heterogeneity both within and between cancer samples. Our group has recently developed methods for cell-level analysis of the multiplexed proteomics image data obtained from colorectal cancer samples. These methods together with the highly informative multiplexed proteomics image data hold great promise for discovering complex biomarkers that can aid the development of personalised medicine.

Introduction
Cancer is continuously revealed to be ever more complex than previously thought. Recent studies using a variety of new proteomics technologies have revealed a surprisingly large degree of variation between individual cancer cells from the same tumour [1,2]. Current clinical practice uses biomarkers limited to the simultaneous analysis of only a handful of proteins. These biomarkers fail to assess the true complexity of cancer, and the resulting biomarkers have a low prognostic value [3]. This may be because tumour cell heterogeneity strongly influences treatment-unresponsiveness and treatment-resistance [4]. Hence, there is a need for novel multiplex markers that simultaneously capture the expression of numerous proteins with relation to each other. Therefore, an imaging system capable of studying heterogeneity in situ is potentially a very useful clinical research tool.

New biomaging techniques have recently been proposed to visualise the co-location or interaction of several proteins in cells in intact tissue specimen. These include MALDI imaging [5], Raman microscopy [6], Toponome Imaging System (TIS) [7], MxF [8] and multi-spectral imaging methods [9]. 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 spatial location on a single tissue section [10]. It runs cycles of fluorescence tagging, imaging and soft bleaching in situ and generates multi-channel image data, where each individual channel provides information about the abundance level of a specific protein molecule. While colocation does not necessarily imply interaction, it is a potential indication for an interaction that is not necessarily direct. The images were acquired using a TIS microscope (ToposNomos Ltd., Munich, Germany), which has a sub-cellular maximum lateral resolution of 206 × 206 nm/pixel [11]. This allows the determination of sub-cellular protein network architectures.

Developing quantitative methods to analyse the large amounts of data generated can aid the development of new sensitive and specific multiplex biomarkers – also termed as complex biomarkers – for risk stratification and diagnosis, as well as anti-cancer drug discovery by identifying ‘hub’ proteins that are essential regulators of protein networks [12].

Methods
Recently, a departure has been made from the pixel-level analysis, commonly employed in the analysis of combinatorial molecular phenotypes (CMPs) [7] or molecular co-expression patterns (MCEPs) [13-15] in TIS imaging data. Examining the image data at a cell level allows us to study the heterogeneity of cell phenotypes within cancer samples. Briefly, the samples used for the recent cell-level analysis frameworks developed by our group at Warwick 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. A library of 12 antibody tags, some of which are known tumour markers or cancer stem cell markers, was used. These were CD133, CK19, Cyclin A, Muc2, CEA, CD166, CD36, CD44, CD57, CK20, Cyclin D1 and EpCAM. The stacks also included a DAPI tag used to identify the cell nuclei. The image stacks obtained are then pre-processed using the following steps. Firstly, the images are aligned using the RAMTaB (Robust Alignment of Multi-Tag Bioimages) algorithm [16]. This is done in order to prevent potentially erroneous co-mapping resulting from the slight mis-alignment of the multi-tag images obtained using TIS. Background autofluorescence is digitally removed so any remaining signal should be true protein expression. The images are then segmented using a modified form of a graph cut based method [17] applied to the DAPI channel [18] (Figure 1). This is necessary in order to extract pixel locations of the nuclei and their immediate neighbourhood only, as the DAPI tag stains the DNA. This provides a rough approximation of the cells and removes possible noise from the stromal and lumen areas.

One of the cell-level analysis frameworks was presented by Khan et al. [20]. This work is an extension of a previous study [18] mining for cell phenotypes based on their high-dimensional protein co-expression profiles. Khan’s study showed that the raw protein expression vectors have a non-linear high-dimensional structure, which can be effectively visualised using a symmetric neighbourhood embedding approach. The three-dimensional vectors obtained were found to outperform the original high-dimensional raw protein expression vectors in terms of their ability to discriminate between normal and cancer tissue samples on the basis of their phenotypic distributions [20].

Another related study proposed a way to analyse phenotypes obtained according to protein-protein dependence profiles of the cells [19]. Here, we proposed a new measure called Difference in Sum of Weighted Co-dependence Profiles (DiSWOP) that can highlight protein pairs that are more co-dependent and more frequently occurring in cancer than in normal samples, or vice versa. The measure weights the dependency score of a protein pair with the phenotype probability in the sample, and sums all occurrences of the protein pair in all the cancerous samples and in all the normal samples. The sums are normalised 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 [19]. Applying this measure to cell phenotypes obtained using Affinity Propagation clustering [21], several protein pairs were highlighted in terms of a small social network of proteins, as shown in Figure 2 [19]. One of these pairs is CEA and EpCAM, which came out as more co-dependent in the cancer samples. This protein pair has also been found experimentally to interact via the pathway CEA–SOX9–Claudin7–EpCAM, which is involved in determining the morphology of the colon epithelium [22-25]. The protein pairs highlighted by the measure are very similar when other phenotyping methods are used. This has been demonstrated by using Bayesian hierarchical clustering with a Gaussian prior [26] on half of the protein pairs, which discriminate best between cancer and normal tissue [25]. In addition, further validation of the measure has been done using synthetically generated image data [19].

Discussion and conclusions
The methods presented above have been applied only on a small number of samples – 3 [20] and 11 [19, 25]. In an on going
References


