IA-2

Automated Detection of Circulating Tumour Cells

C-M. Svensson, S. Krusekopf, J. Lücke, MT Figge

01/09/2013

Leibniz Institute for Natural Product Research and Infection Biology - Hans Knöll Institute

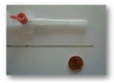
Automated Detection of Circulating Tumour Cells

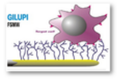
Carl-Magnus Svensson[1], Solveigh Krusekopf[2], Jörg Lücke[3] and Marc Thilo Figge[1,4]
[1]Research Group Applied Systems Biology, HKI and [2]Friedrich Schiller University Jena, Germany
[3]GILUPI GmbH, Potsdam, Germany
[4]Department of Software Engineering and Theoretical Computer Science, TU Berlin, Berlin, Germany

1. Introduction

We have developed a method to identify and count Circulating Tumour Cells (CTCs) in fluorescence microscope images. CTCs are viable cells that detached from primary tumour sites and disseminate via the bloodstream as seeds for secondary tumours in other organs The detection and enumeration of rarely occurring CTCs is an important diagnostic tool in evaluating the progression of disease and the effectiveness of treatment [1].

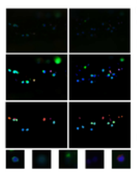
2. Data Collection





The data of this study were generated by a technology that can collect CTCs *in vivo* using a *functionalized and structured medical wire* (FSMW) [2]. The tip of such a wire is functionalized by the application of EpCAM antibodies and is inserted into the cubital vein of a patient. It is left there for 30 minutes during which the antibody-coated tip is collecting CTCs from the blood that flows past. After collection of CTCs, the samples are stained and CTCs are manually enumerated directly on the FSMW using fluorescence microscopy.

3. Data and ROI Identification



The identification of regions of interest (ROIs) are found through the following steps:

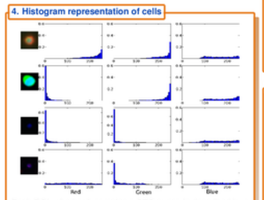
Contrast Normalisation using a Naka-Rushton filter.

Blue Channel Thresholding CTCs are exhibiting colloqualized EpCAM and DAPI staining (blue and green channels) regions of increased blue is found by thresholding

Watershed Segmentation Regions above threshold are decided with using the watershed algorithm. Our dataset consisted of 61 images in which the watershed algorithm found 3545 foreground regions.

Morphological screening Found ROIs are screened based on size and roundness to eliminate regions that obviously are not cells. This left us with 617 ROIs being considered as possible CTCs.

4. Histogram representation of cells



Each ROI, whether it contains a CTC or not, is described by a RGB-histogram based on the colour content of the ROI. The histograms of individual ROIs are compared with class histograms that are learned using maximum likelihood. The difference between histograms of data points, $\boldsymbol{y}_n$, and the class histograms, $\boldsymbol{H}_c$, are measured by quadratic-form divergence $D_A = \sqrt{(\boldsymbol{y}_n - \boldsymbol{H}_c)A(\boldsymbol{y}_n - \boldsymbol{H}_c)}$. The divergence within a class is assumed to follow a half normal distribution

$$p(\boldsymbol{y}_n|c;\boldsymbol{H}_c,\sigma_c) = 2\mathcal{N}(D_A(\boldsymbol{y}_n,\boldsymbol{H}_c);0,\sigma_c),$$

where σ_c denotes the standard deviation of the normal distribution.

5. Generative Model

The classification of cells is made based on the probability function $p(c|\boldsymbol{y}_n)$ which express the probability that the data point, $\boldsymbol{y}_n$, belongs to class c.

$$p(c|\boldsymbol{y}_n) = \frac{p(\boldsymbol{y}_n|c)p(c)}{\sum_{c'} p(\boldsymbol{y}_n|c')p(c')}, \quad p(c|\boldsymbol{\pi}) = \pi_c.$$

The trained generative mixture model can be used for classification of an unknown data point that is presented by calculating the posterior probability using Bayes theorem. The appropriate class is then chosen by taking

$$c^* = \operatorname{argmax}_c(p(c|\boldsymbol{y}_n;\theta)),$$

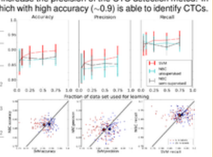
i.e. we assign the data point to the class that maximises its posterior probability. The probabilistic model can be trained unsupervised, supervised or semi-supervised. For unsupervised learning we use the Expectation Maximisation algorithm where we maximize the data log likelihood,

$$\log(\mathcal{L}(\theta)) = \sum_{n=1}^{N} \log(p(\boldsymbol{y}_n;\theta)).$$

6. Results

The accuracy, precision and recall of the individual classifiers on the complete data set. The first row shows the results of the ROI identification step where ROIs that possibly contained CTCs are separated from debris using a decision tree. Furthermore, the results of histogram-based retrieval are shown for the classifiers SVM with RBF-kernel, unsupervised and semi-supervised NBC. The ROI identification process captures almost all the CTCs that we have in the data set (recall being 0.96) but about half of the ROIs do not contain CTCs (precision is 0.51). The identified ROIs are then classified using NBC or SVM which greatly increase the precision of the CTC detection method. In conclusion we have a classifier which with high accuracy (~0.9) is able to identify CTCs.

model	Accuracy	Precision	Recall
ROI identification	0.99	0.51	0.96
SVM RBF-kernel	0.89 ± 0.02	0.87 ± 0.03	0.93 ± 0.02
NBC unsupervised	0.87 ± 0.02	0.85 ± 0.04	0.92 ± 0.03
NBC semi-supervised	0.88 ± 0.02	0.85 ± 0.03	0.93 ± 0.02



References

[1] Banys *et al.*, *Clinica Chimica Acta*. 423:39–45, 2013.
[2] Saucedo-Zeni *et al.*, *International Journal of Oncology*, 41(4):1241–1250, 2012.

Leibniz Association

Contact: carl-magnus.svensson@hki-jena.de