Towards image-based cancer signatures from histopathology data

Annual activity report

Supervisor : T. Walter and F. Reyal.
Units: Center for Computational Biology and UMR900

Peter Naylor

Contact information:
Email: peter.jack.naylor@gmail.com
Tel: *+33 6 72 09 66 74*
Address: 21 rue Bizet
94800 Villejuif (FR)

June 2016

Towards image-based cancer signatures from histopathology data

Peter Naylor
Supervisor: Thomas Walter and Fabien Reyal.

June 29, 2016

The summary of this report is the following. I will start by giving a brief introduction of my PhD subject, continued with the main difficulties to overcome and a brief bibliography. A chronological assessment of my work follows: from November to April I worked on tissue segmentation, Nuclei segmentation is my current focus and I will conclude with my future work.

1 PhD subject: objectives and strategy

This PhD project aims at developing the tools to take advantage of the morphological and spatial information at the cellular and tissular scale from histopathology data.

The basic work flow is shown in Figure 1. Tissue samples are taken from the breast prior to surgery. In parallel, slides are prepared and the tissue is profiled in terms of gene expression and / or sequencing. From the work flow, we aim at extracting physiologically relevant features, which can then be used (optionally in combination with expression and mutational data) to predict either the molecular cancer subtype or the prognosis for the patient.

The main focus of this PhD thesis will be the extraction of physiologically relevant features at the cellular and tissue level. Regarding the *cellular level*, I will focus on nuclear morphologies, because (1) nuclei are indicative of many cellular phenotypes[3] and (2) their morphology is currently used by pathologists in order to identify the mitotic index and the level of nuclear pleomorphism[4]. In order to derive information at the cellular level, nuclei must first be identified by automatic segmentation. Second, we can design cellular classifiers that assign to each of the segmented nuclei a cell type labeled (such as epithelial cell, stromal cell or lymphocyte) or a cellular phenotype (such as cell death, interphase, metaphase). This can be achieved by supervised machine learning approaches, such as Support Vector Machines (SVM) or Random Forests (RF), where the classification rules are infered in a fully automatic way from annotated samples. Some of the features will also be exported directly, as they are themselves physiologically relevant (such as cell size). Regarding the *tissue level*, the plan is to detect the tumor, stromal and necrotic regions (regions containing mostly dead cells). The tissue level features I will consider, are on the one hand region based features, calculated on the regions, and on the other hand features calculated from the cell populations (such as cell type percentages, and features describing the level of organization, such as Ripley's K), stratified by the regions in which the cells are situated. In both cases, segmentation will be probably a bottleneck, and I am particularly interested by methods that are easily adaptable to new data sets and new segmentation tasks.

Data sets I will apply these methods to two datasets: (1) 208 slides from an unpublished study on breast cancer, a special type of very aggressive breast cancer and (2) 198 slides from a recently published study on bladder cancer [1]. In the first data set, I will be able to study the predictability of treatment response by automatic analysis of histopathology data. The second data set will be informative about how the histopathology features correlate with the molecularly defined subgroups. Indeed, I hope to identify links between cellular phenotypes, transcriptomic and grading data that will feed future projects in this field with interesting hypotheses.

In this PhD thesis, I want to contribute to the generation of the appropriate tools to quantify the huge amount of data found in histopathology slides. On the long run, such a quantification scheme would fit in a work pipeline that would investigate the most informative physiological features at the cellular and the tissue level and the links to genomic, transcriptomic features and even possibly different medical imaging such as 3D MRI scans.

2 Context and difficulties

So far, histopathology data is still largely unexploited in a systematic and quantitative way. There are several reasons for this:

1. With the availability of comprehensive genome, transcriptome and epigenome data sets, the hope was that these data could explain many aspects of life and virtually all aspects of diseases which are known to be related to the genome. Today, we understand that it is necessary to include the spatial and morphological dimensions in the reasoning.

2. Digital pathology, has only recently become a standard in the field. Still a few years ago, the standard procedure was to examine the slides on the microscope. We now have powerful scanners that produce whole slide images.

3. The information contained in histopathology data has never been fully formalized. While some quantitative (yet manually determined) criteria exist, this is not true for the overall interpretation of slides which remains subjective and difficult to model formally. This issue is also linked with the high variability between slides, which is due to the stain variability between and the tissue variability.

4. There are many technical and methodological challenges related to the automatic analysis of histopathology data which made fully automated analysis of these images unfeasible. In particular, we can think of the size of the data, more than $50GB$ uncompressed data per slide and we have hundreds of slides.

However, many people have started to work on histopathology slides, these works try to cope with these difficulties and are often focused on the detection of objects within histopathology slides. In some recent publications, the correlations to clinical variables have also been investigated, as well as their capacity to complement other types of data, such as expression data. Image-based biological features are also emerging: In Lee et al. [7], they improve the prediction of recurrent prostate cancer via the integration of quantitative image features and protein expression. Importantly, histopathology data gives access to the single cell level and allows to evaluate the heterogeneous expression of biomarkers. In Potts et al. [11], the authors analyze both intratumoral heterogeneity (within a single tumor) and intertumoral heterogeneity (between tumors at different sites). In particular, they enrich HER2 scoring schemes in order to take heterogeneity into account, via features based on spatial distribution and local neighbourhoods. Harder et al. [5] find relevant biological features that quantify the invasion of the tumor, in particular these features are based on spatial and neighbourhood analysis. In Petushi et al. [10], they show that relevant data can be extracted from the images in order to create reproducible grading. Reproducibility of the grading is a key goal for computer aided diagnostics, indeed breast cancer grading and many others can be highly variable from one pathologist to another, and even more so from one hospital to another.

3 Tissue segmentation

Histopathology annotated data is very scarce, indeed, whole slide images (WSI) are very large files that need expert pathologist in order to highlight the relevant information. Very recently however, a large data set of manually segmented regions has been made available to the scientific community (Camelyon16, published in the framework of ISBI 2016). This data set contains 270 annotated WSI provided by two different hospitals, every metastasis region in this data set is annotated, see figure 4.

My first approach to segment these image data was based on a supervised learning workflow, where I defined a set of features and tested several algorithms for automatic classification. These hand crafted features were based on two distinct methods: some were the result of the application of image filters at different scales, gaussian smoothing, Laplacian of Gaussian, etc. [13]. Others were extracted via the use of mathematical morphology and in particular waterpixels, [9]. The strategy was thus to first partition the image into small homogeneous regions for which texture and intensity features can then be calculated. The workflow for the feature extraction can be found in figure 5. Once we had the features we applied a random forest tuned via cross validation. The training was a delicate part as the data set was huge and highly unbalanced, more than 5.10^{12} lines and only 1% were positive. We used a modified random forest in order to speed up convergence, the sampling procedure was modified in order to reduce the computational burden. We presented a unique part of the data set to each tree construction instead of the whole dataset, see [2]. This novel random forest wasn't any less accurate then the standard random forest, however the speed up was considerable. From this classifier we were able to produce posterior probability maps, see figure 6.

While this first approach gave overall reasonable results, the posterior probability maps were also relatively noisy and could not cope very well with the highly variable metastatic regions. There are two explanations for this: either the subsampling was too restrictive and the amount of data presented to the classifier was too small as compared to the difficulty of the task, or the hand crafted features were not adapted for this task. These hypotheses suggest the use of deep networks and in particular deep convolutional network such as ImageNet [6] and GoogleNet [14]. I am therefore currently working on the setup of this methodology (both hardware and software) in order to segment both nuclei and regions in histopathology slides.

4 Nuclei segmentation

Nuclei segmentation annotation for histopathology is an even scarcer dataset, the number of cells in one histopathology slide could reach millions of cells. The workflow for this part is pictured in figure 2. In order to apply supervised classifiers we started creating our own manual annotation for a limited number of patches of the slides. In order to help the workload, we used a previous unsupervised segmentation approach to do a first segmentation, if this first segmentation is too poor we discard it, if not we correct it and save the patch. The software used for the manual segmentation is *itksnap*. Once we have sufficient annotation samples we will be able to train, or *fine-tune* our network. We will be using fully convolutional network for their state of the art semantic segmentation on Image net [8] and in medical imaging, [12]. Once the segmentation over the whole slide is achieved, we can consider extracting object specific features in order to classify each nucleus with respect to his type (lymphocyte, epithelial cell, cancerous cell and so on) but also to his current state (undergoing mitosis, apoptosis and so on). Thank to this classification we can create density maps over the slide and create other biological relevant features, such as lymphocyte density counts and maybe spatial distribution features. The relevance of such features have been shown in recent works [15, 7].

Once the nuclei segmentation step achieved, we will be able to use the information provided by the density counts in order to derive more information about the tissues, for instance detecting metastatic regions, necrotic tissue and healthy tissue. All the information gathered on one WSI will be summed up as a highly dimensional feature vector that will try and describe the histopathology data in a biologically relevant aspect for one patient.

5 Future work

Once the computer vision aspect concluded, which is one of the bottlenecks for this PhD subject, I will focus on the different aspect present in figure 1. In particular extracting features from the genomic data and combining the very heterogeneous data set is a difficult task. Indeed, even if the extraction procedure are independent, the data will stay highly dependent. We wish to explore links between genomic signatures and phenotypic patterns, exploring the links between these heterogeneous type of data will enable us to better exploit the present complementary information between them in order to increase the performance of only genomic models. Other works have investigated the combination of such heterogeneous type of data. As an example, Yuan et al. [15] segment nuclei in histological data to access information about

lymphocyte counts, cancer cells counts and heat maps of spatial distribution in order to improve genomic based models. They also use these biological driven features to correct statistical models based on molecular data, indeed they assess that these biologic driven features quantify the heterogeneous cell populations that is a source of noise.

References

[1] Anne Biton, Isabelle Bernard-Pierrot, Yinjun Lou, Clémentine Krucker, Elodie Chapeaublanc, Carlota Rubio-Pérez, Nuria López-Bigas, Aurélie Kamoun, Yann Neuzillet, Pierre Gestraud, et al. Independent component analysis uncovers the landscape of the bladder tumor transcriptome and reveals insights into luminal and basal subtypes. *Cell reports*, 9(4):1235–1245, 2014.

[2] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[3] Kin-Hoe Chow, Rachel E Factor, and Katharine S Ullman. The nuclear envelope environment and its cancer connections. *Nature reviews. Cancer*, 12(3):196–209, mar 2012. ISSN 1474-1768. doi: 10.1038/nrc3219. URL http://www.ncbi.nlm.nih.gov/pubmed/22337151.

[4] C W Elston and O Ellis. Pathological prognostic factors in breast cancer . I . The value of histological grade in breast cancer : experience from a large study with long-term follow-up. *Histopathology*, 19:403–410, 1991.

[5] Nathalie Harder, Maria Athelogou, Harald Hessel, Alexander Buchner, Ralf Schüonmeyer, Günter Schmidt, Christian Stief, Thomas Kirchner, and Gerd Binnig. Co-occurrence features characterizing gland distribution patters as new prognostic markers in prostate cancer whole-slide iamges. In *International Symposium on Biomedical Imaging (ISBI'16), Prague, Czech Republic, April 13-16, 2016*, 2016.

[6] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[7] George Lee, Asha Singanamalli, Haibo Wang, Michael D Feldman, Stephen R Master, Natalie NC Shih, Elaine Spangler, Timothy Rebbeck, John E Tomaszewski, and Anant Madabhushi. Supervised multi-view canonical correlation analysis (smvcca): integrating histologic and proteomic features for predicting recurrent prostate cancer. *IEEE transactions on medical imaging*, 34(1):284–297, 2015.

[8] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3431–3440, 2015.

[9] Vaïa Machairas, Matthieu Faessel, David Cárdenas-Peña, Théodore Chabardes, Thomas Walter, and Etienne Decencière. Waterpixels. *Image Processing, IEEE Transactions on*, 24(11):3707–3716, 2015.

[10] Sokol Petushi, Fernando U Garcia, Marian M Haber, Constantine Katsinis, and Aydin Tozeren. Large-scale computations on histology images reveal grade-differentiating parameters for breast cancer. *BMC medical imaging*, 6(1):1, 2006.

[11] Steven J Potts, Joseph S Krueger, Nicholas D Landis, David A Eberhard, G David Young, Steven C Schmechel, and Holger Lange. Evaluating tumor heterogeneity in immunohistochemistry-stained breast cancer tissue. *Laboratory Investigation*, 92(9):1342–1357, 2012.

[12] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015*, pages 234–241. Springer, 2015.

[13] Christoph Sommer, Christoph Straehle, Ullrich Koethe, and Fred A Hamprecht. ilastik: Interactive learning and segmentation toolkit. In *Biomedical Imaging: From Nano to Macro, 2011 IEEE International Symposium on*, pages 230–233. IEEE, 2011.

[14] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–9, 2015.

[15] Yinyin Yuan, Henrik Failmezger, Oscar M Rueda, H Raza Ali, Stefan Gräf, Suet-Feung Chin, Roland F Schwarz, Christina Curtis, Mark J Dunning, Helen Bardwell, et al. Quantitative image analysis of cellular heterogeneity in breast tumors complements genomic profiling. *Science translational medicine*, 4(157):157ra143–157ra143, 2012.

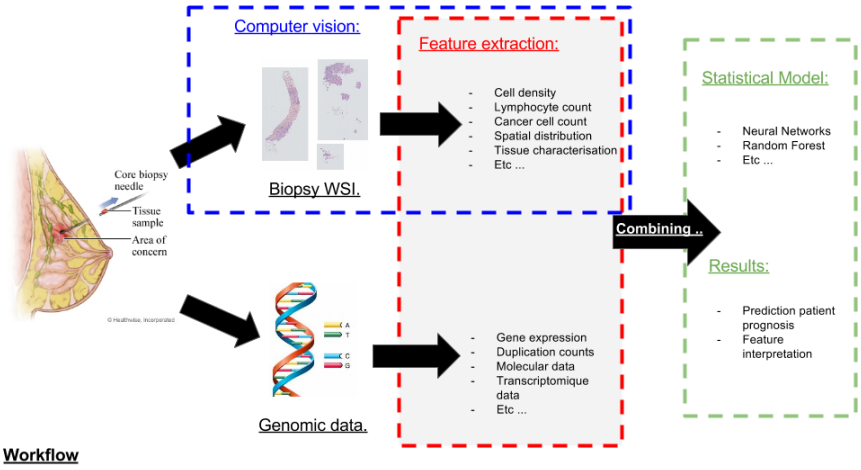


Figure 1: Workflow

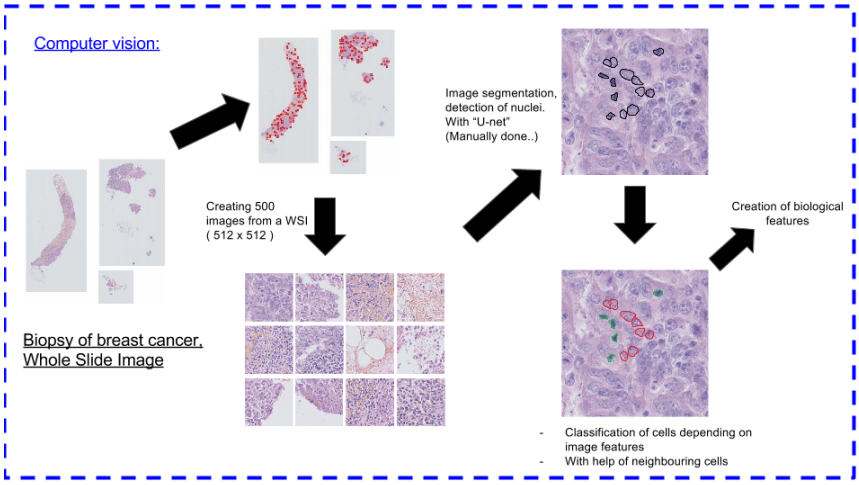
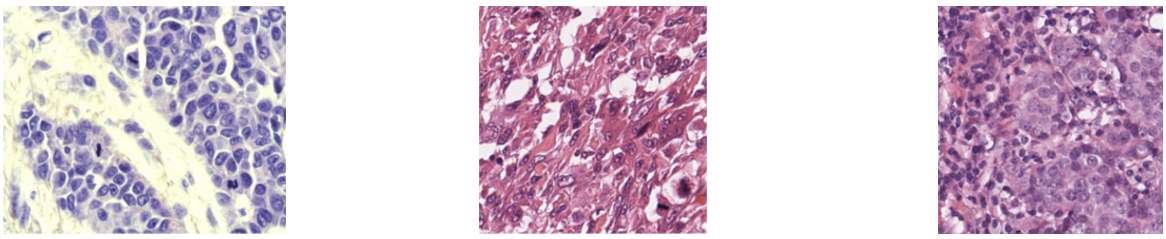


Figure 2: Computer Vision aspect



(a) (b) (c)

Figure 3: Tissue sections with standard Haematoxylin and Eosin staining.

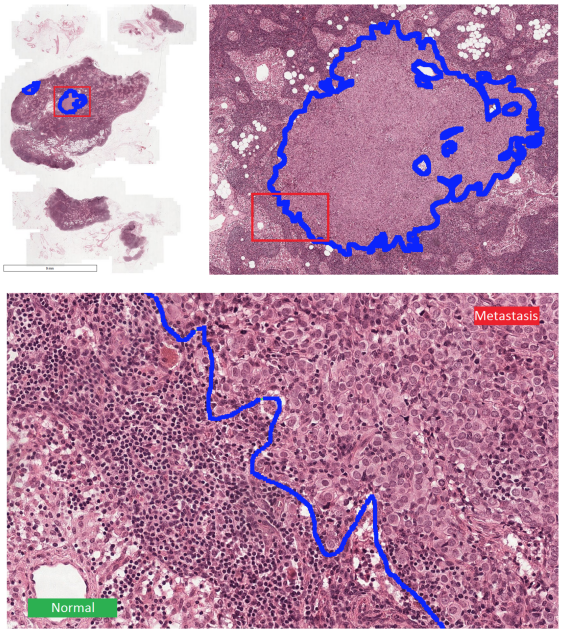


Figure 4: Annotated data

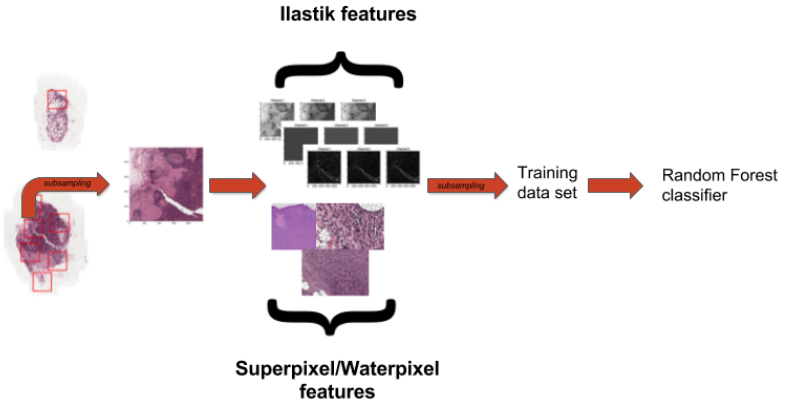


Figure 5: Work flow for metastatic region detection

(a) RGB raw data. (b) Probability map. (c) Metastasis confidence pin-points.

For figure (c), blue is equal to a low confidence score whereas red means a high confidence.

Figure 6: Evaluation whole slide image, test slide number 2.