Towards the identification of tissue-based proxy biomarkers

Vlad Popovici, Ph.D.[1]
[1]**Institute of Biostatistics and Analyses, Masaryk University, Brno, Czech Republic**

Abstract

Accurate patient population stratification is a key requirement for a personalized medicine and more precise biomarkers are expected to be obtained by better exploiting the available data. We introduce a novel computational framework that exploits both the information from gene expression data and histopathology images for constructing a tissue-based biomarker, which can be used for identifying a high-risk patient population. Its utility is demonstrated in the context of colorectal cancer data and we show that the resulting biomarker can be used as a proxy for a prognostic gene expression signature. These results are important for both the computational discovery of new biomarkers and clinical practice, as they demonstrate a possible approach for multimodal biomedical data mining and since the new tissue-based biomarker could easily be implemented in the routine pathology practice.

Introduction

Tumor heterogeneity is a major hurdle on the path to a personalized medicine and many molecular biomarkers have been proposed for stratifying the patient population and predicting the (non) response to treatments. This heterogeneity is obvious both at tissue level, in histopathology images, and at molecular level, with various gene and protein profiling studies proposing systems for subtyping the cancers. With the recent technological developments, it is now possible to interrogate the same underlying biological reality from different perspectives and at various resolutions, from sequencing of the whole genome of an individual cell to magnetic resonance imaging of an organism. The combination of these diverse modalities is not only natural, but also necessary if we are to advance our understanding of the biological processes and to fully exploit the available data. Some combinations are more amenable to a common representation, thus facilitating the interpretation and inference – for example methylation and expression data or computer tomography and magnetic resonance imaging. In other situations, it is not clear if, how and to what extent some modalities can be combined. This is the case for microscopy imaging and gene expression (or other similar molecular data) – a combination of primary importance since it would allow establishing direct genotype-phenotype correlations, where tissue architecture and heterogeneity could be linked to some molecular signals. In the particular case of oncology, such correspondence may open new directions for identifying gene responsible for certain pathological phenotypes or, the primary goal of the present study, for identifying tissue biomarkers that can be used as proxies for genomic signatures.

Some recent results indicate the potential of this approach. For example, in the case of breast cancer a few genomic-phenotypic correlations are known, such as the association of lobular phenotype with deletions in the CDH1 gene (encoding for E-cadherin)[1] or the predictive utility of mesenchymal/metaplastic features in the case of AR-positive triple negative breast cancers[2]. In the case of colorectal cancer (CRC) it is well known the association of mucinous/serrated carcinomas with BRAF mutations and we have shown that such association can be extended to a larger group of "BRAF mutated-like" tumors, characterized by a specific genomic signature[3]. Similarly, connections between nuclear morphometry and molecular data have been identified in glioblastoma[4,5] and exploited in a multimodal prognostic signature in breast cancer[6]. These observations all support the idea that genomic and phenotypic traits can be put in correspondence and, by consequence, that some phenotypic features could potentially be used as proxies for genomic markers.

In the present work we propose a computational framework for jointly mining the genomic and imaging data. The gene expression data is supposed to be obtained from the same (or adjacent) tumor section as the histopathology whole-slide image. The key point of our approach resides in a convenient re-coding of the imaging data that makes it usable by the standard data mining methods. In usual tissue-based gene expression profiling experiments, the DNA (or RNA) is extracted from a pool of cells, either micro- or macro-dissected from a designated tissue region. In the end, the value associated with a gene (or probeset) and considered to represent its expression level can be seen as an average signal over all the extracted region/cells[*]. On the other hand, a whole-slide pathology image has a number of pixels in the order of $10^9 - 10^{10}$. Clearly, using raw pixel values (color or simply intensity) to find

[*] Single cell profiling or sequencing is possible[7], but not yet feasible for every cell in a whole region of interest and the utility of such an approach for large studies remains to be demonstrated.

connections with the molecular data is nonsensical since isolated pixels do not bear enough information. Thus we are looking for a representation of image data in terms of features that capture local aspects of the image, ideally in a multi-scale and rotation invariant fashion, and which can be aggregated to give a global characterization of the tissue region of interest. This is a typical problem in content-based image retrieval applications and, inspired by them, we adapted here one of solutions proposed in this context – the *bag-of-features* method.

To demonstrate the feasibility of the proposed framework, we explore whether the image-based features can be used – and to what extent – to predict a classification based on a genomic signature, namely the BRAF mutant-like (BL, hereinafter) signature[3]. This gene expression signature has been shown to identify a high-risk group of CRC patients[3] constituting a well-defined subpopulation[8] for which targeted therapies could eventually be developed[9]. The group of BL tumors (which includes about a third of all KRAS mutants), while being BRAF V600E wild-type, share a number of clinicopathologic characteristics with the true mutants, including poor overall and after relapse survival, higher frequencies of mucinous histology, high grade tumors, and microsatellite instability. The gold standard for identifying this subpopulation is given by the genomic signature, which involves computing a score as a linear combination of 64 genes[3] (a difference of means of two groups of genes), with a positive score indicating a BL sample and a negative score a non-BL sample, respectively. The BL patient population in described in *Popovici et al.*[3].

Methods

An overview of the computational framework is given in Figure 1. There are three distinct phases: learning a suitable image representation (codebook learning), image re-coding and multimodal data mining/analysis. Once the image representation is fixed and all images are recoded, we learn a binary classifier to predict gene expression-based labels. All modeling (learning the image representation and the classifier) is performed on a modeling set,

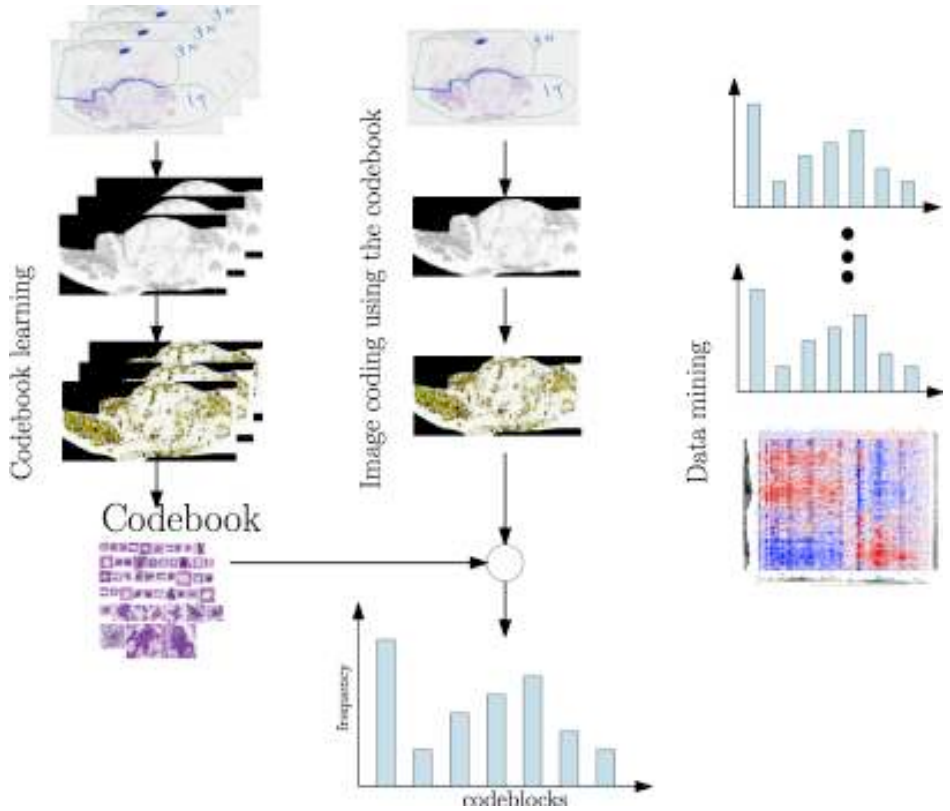

Figure 1. A schematic overview of the main steps of the computational framework: From a set of training images the corresponding tumor regions are extracted (following the annotation) and the local descriptors are computed (colored regions in the image). Using *k*-means clustering, a codebook is constructed. A new image follows the same image preprocessing steps up to local descriptor extraction. Then, based on the codebook, the local descriptors are assigned to codeblocks whose occurrence frequencies are saved as the image coding. With the images recoded, joint image and molecular data mining may proceed as usual.

independent on the validation set. We discuss the details of this framework and the statistical methods used for analysis in the following sections.

Data

The data collection used consisted of 113 samples for which both histopathology whole-slide images and clinical data were available, along with corresponding BL genomic scores[3] (computed from gene expression data measured from a nearby tumor section). These samples are a subset of those described in *Popovici et al.*[3].

All samples were BRAF V600E wild-type, stage III, microsatellite-stable, CRC tumors. The collection was split in two sets for modeling ($n_{tr} = 40$) and testing ($n_{ts} = 73$). In the modeling set the two groups of BL (positive BL score) and non-BL (nBL – negative BL score) were equally sized, while in the validation set 44% were BL.

For each sample, a whole-slide scan was obtained at $40\times$ magnification (Hamamatsu scanner, typically of the order of $150{,}000\times100{,}000$ pixels) and, for practical reasons, was computationally downscaled to an equivalent $2.5\times$ magnification.

Image preprocessing

Since the gene expression was measured only from the tumor regions in each section, the image analysis was confined to the same regions (marked in the slides by an expert pathologist before DNA extraction). Hence, the first preprocessing step was to crop down the images to the bounding box of the tumor regions and to remove the normal tissue region, followed by a denoising step (Gaussian filtering). The images represented standard pathology slides, stained with haematoxylin (binding to chromatin-rich regions such as nuclei) and eosin (as contrast), from which only the haematoxylin was used in subsequent analyses. The intensity of the haematoxylin signal was obtained by color deconvolution[10].

Feature extraction and image re-coding

As discussed in the Introduction section, the gene expression "signal" can be seen as an average over all cells from the micro-dissected tissue region (tumors, in our case). The chosen image representation is somehow similar: the *bag-of-features*[11] represents the image in terms of frequencies of some basic image blocks. This method produces a sparse representation of an image and is an adaptation of the bag-of-words technique from text categorization and retrieval applications. It relies on the construction of a representative codebook, followed by the re-coding of the images in terms of "words" (codeblocks) from this codebook[11]. Typically, the construction of the codebook begins with the computation of a number of feature vectors that represent some characteristics of the images over small local neighborhoods. The selection of these neighborhoods may be systematic (all the patches in a given region of interest), random or may rely on some points-of-interest detector. Once a representative collection of feature vectors has been obtained, a codebook is constructed by $k-$means clustering of the vectors, with k being a user-supplied parameter. The resulting k cluster centroids (called codeblocks), $C_i, i = 0, \ldots, k-1$, form the basis of the image re-coding: a given image is re-coded by assigning all extracted local neighborhoods to one of the k clusters (nearest centroid classification) and building a vector of codeblock occurrence frequencies (Figure 2). Thus, each image is represented in terms of a vector from $[0,1]^k$. Once the codebook has been constructed, for any new image a similar k-value vector is obtained by following the same steps as before, with the exception that the codebook is fixed. This is a standard vector quantization method in signal and image processing, used initially for signal compression[12]. The choice of clustering algorithm ($k-$means) is justified by its probability density modeling properties and its scalability to very large data sets, as is the case in image processing.

In the present work, this general framework has been modified in two key aspects: (i) we introduce a codebook size optimization step and (ii) we allow for uncertainties in image recoding. The codebook size optimization is aimed at improving the image recoding with respect to the binary classification problem to be solved. The small sample size ($n_{tr} = 40$) drastically reduces the design choices and possibilities of parameter optimization, while increasing the chances of overfitting. We have chosen to use a diagonal linear discriminant (DLDA) built on the top 10 (the size of the smallest codebook) features selected by two sample t-test, with the side note that a larger modeling set would allow a finer model selection and parameter optimization. The size of the codebook (k) was selected to maximize

the classification accuracy of this model, estimated by 5-fold cross validation (over the training set). Basically, this implied repeatedly constructing a number of codebooks (with $k \in \{10, 20, \ldots, 120\}$), re-coding the images and estimating the performance of the classifier trained to distinguish BL from nBL.

The second modification we introduced is based on the observation that the selection of training images might not cover the whole variability necessary to build a representative codebook. Thus, when re-coding images in the test set, we allowed a certain degree of uncertainty in the codebook by introducing a pseudo-cluster (labeled C_{-1}) to which all local neighborhoods that were too far from any centroid of the codebook were assigned. When constructing the codebook, the average distances μ_i and standard deviations σ_i were recorded for each cluster C_i. Then, for a new test image, any local neighborhood further than $\mu_i + 3\sigma_i$ from any centroid C_i was assigned to cluster C_{-1}. This pseudo-cluster was not part of the model, it merely indicated discarded/not assigned regions. We also tested the classical approach, with no pseudo-cluster, and we assessed the impact of using this strategy on the classification results.

For the extraction of local neighborhoods image descriptors we opted for using Speeded-Up Robust Features (SURF)[13] technique, which combines a key point detector with a local image descriptor to achieve a fast multi-resolution detection and characterization of regions of interest. The descriptor is both rotation and scale invariant, both of which are required properties in our application. We used the implementation provided by the OpenCV library[†] and we set the sensitivity threshold for the detection of key points (Hessian threshold) to 7500 throughout all the experiments. The resulting local descriptor vectors had 64 real values and described local neighborhoods of various sizes (since SURF is adapting to the local image content).

Image classification

Once the codebook size has been fixed and the final version obtained from the modeling set, all images have been re-coded as a set of n_{tr} k-vectors. These vectors were used to construct a DLDA-based classifier to distinguish between BL and nBL classes. Prior to training the classifier, the data was normalized to zero mean and unit standard deviation, in each of the k variables, and normalization parameters saved as part of the model. A simple, t-test based feature ranking procedure was used and only the top l variables retained in the model. In contrast to codebook learning, when the number of variables was kept fixed, here we optimized l over the training set by estimating (5-fold cross-validation) the performance for $l = 10, 15, \ldots, 50$, and selecting the value yielding the highest accuracy.

Statistical analyses

To test for association between the image features and BRAF score we used a two-sided test based on Pearson's product moment correlation coefficient. To test the association between image features and binary variables (BRAF mutant-like and mucinous status) we used two-sided t-test. The 95% confidence intervals (CI) for the classification accuracy were obtained using Agresti-Coull method[14]. Finally, to test the differences in survival between two groups of patients, we used the Gehan-Breslow generalized Wilcoxon test for crossing survival curves, with a significance level of 0.05. No adjustment for multiple testing was applied. All statistical analyses were carried out in `R` statistical computing environment version 3.2.2 with `survival` package 2.38-3.

[†] Available from http://www.opencv.org

Results

Codebook

Using the described procedure, we established that a codebook with $k = 50$ was optimal for our training set. The codebook contained feature vectors corresponding to local neighborhoods of sizes varying from 14×14 to 54×54. The original local neighborhoods (before color deconvolution) are shown in Figure 2.

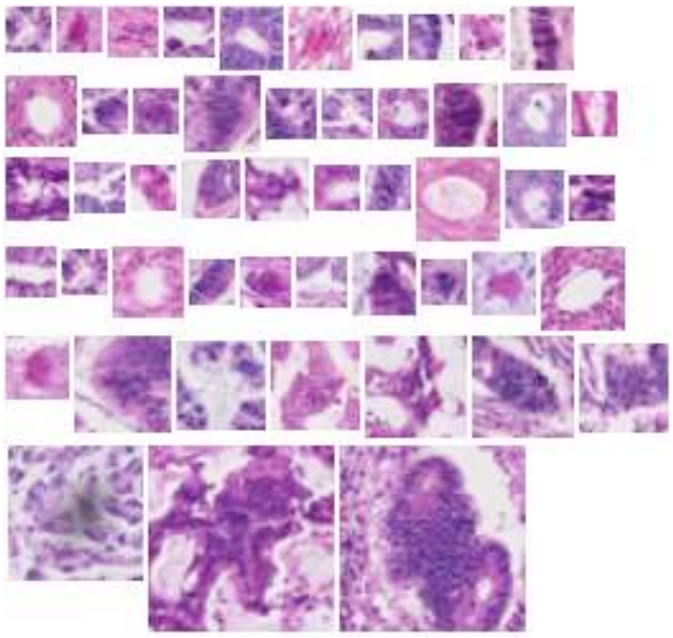

Figure 2. The 50 local neighborhoods whose feature vectors were selected as centroids (codeblocks) of the codebook.

Correlations between image features and genomic score

Each image from the testing set was re-coded in terms of the frequencies of the codeblocks, resulting in a vector of 50 variables. Of these 50 variables, 9 were significantly associated with the BRAF mutant-like status and score (t-test and correlation test p-values < 0.05, respectively) with one additional variable being marginally associated with the score only. Not surprising, the same 10 variables were associated with the mucinous status. No association was found between the pseudo-variable corresponding the cluster C_{-1} and the tested clinical variables.

These initial results showed that the codebook was able to capture, at least partially, the essential features for identifying BL tumors. However, they also showed that none of the feature was strong enough to act as a BL-marker alone, justifying the plan for building a multivariate classifier.

Tissue-based proxy biomarker

We built a DLDA-based classifier to distinguish between the BL and nBL classes, on the re-coded images based on the 50-codeblock codebook. The estimated (by cross-validation, on the modeling set) optimal number of variables was $l = 30$ and led to a model with an accuracy of 91.78% ($95\%\ CI = 82.89 - 96.49$) (sensitivity: 93.75%, specificity: 90.24%) on the independent test set (6 misclassified samples out of 73). In order to assess the quality of the strategy for estimating the optimal number of variables in the classifier (l), several alternative classifiers were built on the modeling set, with the number of variables varying between 10 and 50 (with increments of 5). They have been used to classify the test images that were coded with and without the use of the pseudo-cluster C_{-1}. It was apparent (see Figure 3) that employing a pseudo-cluster led to marginally improved classification accuracy, because the assignment to the 50 codeblocks was stricter. A larger test set would be required to eventually achieve a statistically significant difference between the two scenarios.

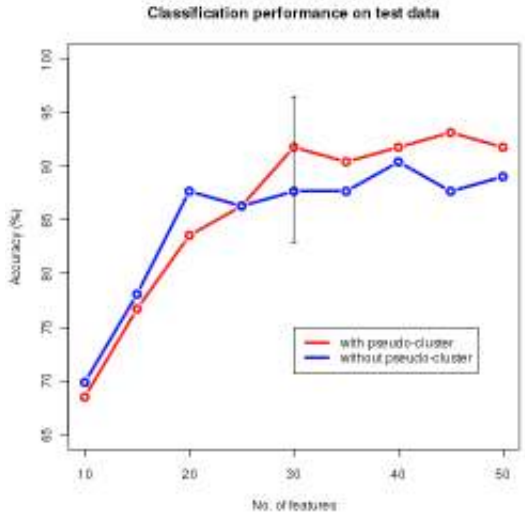

Figure 3. Classification accuracy on the test set for two series of DLDA classifiers: built on codebooks with (red) and without (blue) the pseudo-cluster C_{-1}. For clarity, the confidence interval is shown only for the model selected based on the modeling data.

Population stratification by tissue and genomic biomarkers

The genomic biomarker has been shown to be a strong prognostic indicator for survival after relapse, hence it was hoped that the tissue-based biomarker would have the same property. The tissue-based biomarker misclassified 6 patients (with respect to the genomic signature) and this had a negative impact on the stratification (Figure 4). The stratification obtained by genomic biomarker was statistically significant (p-value=0.02; HR=2.22), while the stratification by tissue-based biomarker showed only a trend towards significance (p-value=0.06; HR=1.62). The image-based classifier had a better sensitivity than specificity so most of the higher risk patients were similarly predicted by both biomarkers (30 out of 32).

On the other hand, the image-based classifier could have been optimized to improve its sensitivity, such that most (or all) of the high risk BL patients to be correctly identified. Such an optimization, while feasible, would require a larger cohort with more events.

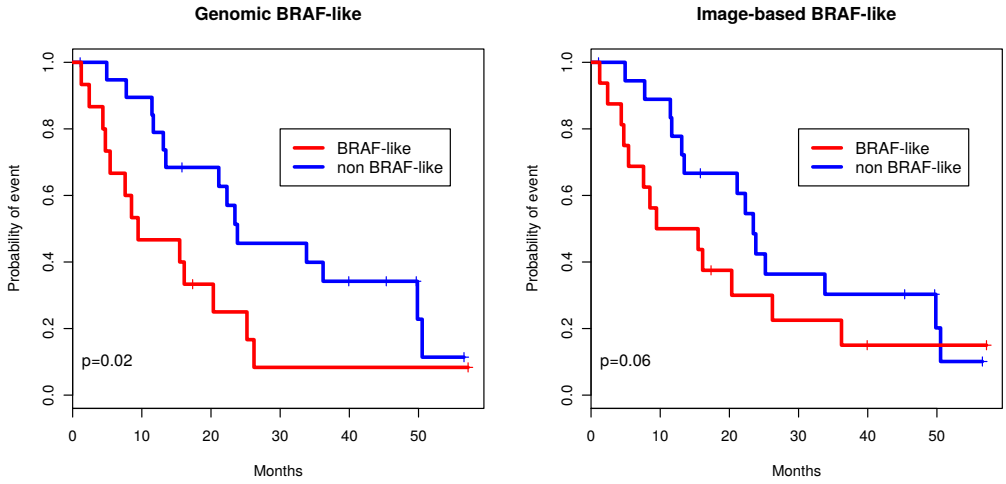

Figure 4. Population stratification by genomic (left) and image-based (right) biomarkers. The 6 misclassified patients have a negative impact on the stratification by the image-based biomarker, but still the stratification remains close to statistical significance (for the a priori chosen significance level).

Discussion

Combining gene expression (or other molecular) data with imaging data may lead to a different perspective on the biological processes. The imaging dimension brings information about phenotypic properties of the sample, whereas the molecular machinery is partially captured by the gene expression data. The main problem in combining such data stems from the difficulty to summarize the images into a reasonably sized set of meaningful (for the task at hand) variables. Indeed, for image data there are a myriad of characteristics that could be measured. One option is to use expert guidance for defining the feature of interest (e.g. nuclear morphometry) and then to build methods able to extract it. The advantage is that such a feature is justified by prior biological knowledge, leading to a model that is easier to interpret. Unfortunately, these features usually require expert-labeled images, which can be quite difficult to obtain and subject to inter-expert variability. Also, being based on previously established results, this approach may limit the ability of making new discoveries. A second option for summarizing the images is using statistics on local descriptors that could be linked in a later stage to some underlying biology. This approach is not confined to prior biological knowledge and may be applied on images that were not previously segmented by an expert. However, it still relies on the data analyst to define the proper local descriptors and is a high risk-high gain approach: there is a higher risk of failure, but also higher chance to establish novel connections.

The work presented here took the second approach and relied on robust local image features that were scale and rotation invariant, making them ideal candidates. Once the images were re-coded based on a codebook adapted to the data, we proceeded to establish connections with gene expression data – here a genomic signature. Several image-based variables were found to be correlated with the "BRAF-mutated-like" status and formed the basis of an image- (tissue-) based biomarker. We call this a "proxy biomarker" since it allows to mimic (to a certain degree) the predictions that the genomic biomarker would make.

The error rate of the image-based biomarker, while encouraging, indicates that there is room for improvement. For example, an improved sensitivity would make the biomarker an interesting screening test, since the patients would be prioritized for genomic testing as well (with strong indications of a possible deregulation of EGFR and/or MAPK pathways[8]).

Our approach used the bag-of-features technique with a modification: in the test images, we allowed for some to be assigned an "unknown" code (denoted C_{-1}) indicating that they were too far from any examples seen in the modeling step. The test results indicated that there is a marginal gain in using this approach (within the confidence interval, for the relatively small test set), which comes with basically no extra costs, thus being worthwhile.

Having such a tissue-based biomarker has direct practical consequences. First, it can lead to a test that can easily be applied, without interfering with the routine pathology workflow, since it only relies on standard images. The pathologist can use the test as an additional source of information, complementing the standard procedures. Alternatively, the test can be applied retrospectively on archive pathology images when searching for cases for expression profiling.

One direction that is not pursued here pertains to the analysis of the identified image features from a pathologist perspective. It would be interesting to see if any of the computationally derived image features can be linked to known morpho-pathology features, validating the approach from a biological interpretation perspective as well. This is left for a follow-up work.

Conclusions

We have presented a novel framework for building tissue-based biomarkers starting from a genomic signature and we applied it to the problem of finding a biomarker able to predict the "BRAF mutated-like"[4] status. This problem has a clear practical importance in the diagnostic and prognostic of colorectal cancer. The samples on which we applied our method were relatively homogeneous (all microsatellite-stable, stage III) hence having a biomarker able to stratify this population is of clinical interest.

We have demonstrated that it is possible to extract, in a data-driven manner, meaningful image features that may be related to genomic data. While this approach would probably not work in all scenarios, it opens the avenue for large-scale automatic data mining, which could lead to the identification of new biomarkers (either tissue-based or multimodal).

As a final remark, the proxy biomarker presented here represents but a first step towards a full-fledged test of clinical relevance. Many more tests and larger sample sizes are required before it would become useful.

Acknowledgements

This project is financed from the SoMoPro II programme. The research leading to this result has acquired a financial grant from the People Programme (Marie Curie action) of the Seventh Framework Programme of EU according to the REA Grant Agreement No.291782. The research is further co-financed by the South Moravian Region. This article reflects only the author's views and the Union is not liable for any use that may be made of the information contained therein.

The author also thanks PETACC3 translational research group and Prof. Fred Bosman in particular, for making available the data for the present study.

References

1. Berx G, Cleton-Jansen AM, Nollet F, et al. E-cadherin is a tumour/invasion suppressor gene mutated in human lobular breast cancers. The EMBO Journal 1995; 14:6-107-115.
2. Lehmann BD, Bauer JA, Chen X, et al. Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies. J Clin Invest 2001; 121:2750-2767.
3. Popovici V, Budinská E, Tejpar S, et al. Identification of a poor-prognosis BRAF-mutant-like population of patients with colon cancer. J Clin Oncol. 2012; 30(12):1288-1295.
4. Cooper LA, Kong J, Gutman DA, et al. An integrative approach for in silico glioma research. IEEE Trans Biomed Eng. 2010, 57(10):2617-21.
5. Kong J, Cooper LA, Wang F, et al. Integrative, multimodal analysis of glioblastoma using TCGA molecular data, pathology images, and clinical outcomes. IEEE Trans Biomed Eng. 2011, 58(12):3469-74.
6. Yuan Y, Failmezger H, Rueda OM, et al. Quantitative image analysis of cellular heterogeneity in breast tumors complements genomic profiling. Sci Trans Med. 2012, 4(157):157ra143

7. Gerlinger M, Rowan AJ, Horswell S, et al. Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. New Engl J Med. 2012; 366(10):883-892.
8. Vecchione L. BRAF mutant colorectal cancer: a different entity. ESMO 16th World Congress on Gastrointestinal Cancer. 2014.
9. San Lucas FA, Fowler J, Kopetz S, Scheet P, Vilar E. Discovering new targeted therapies for BRAF mutant-like colorectal cancers. ASCO Annual Meeting. J Clin Oncol. 2013; 31(suppl; abstr 3623)
10. Ruifrok AC, Johnston DA. Quantification of histochemical staining by color deconvolution. Anal Quant Cytol. 2001, 23(4):291–299.
11. Csurka G, Dance C, Fan L. Visual categorization with bags of keypoints. ECCV International Workshop on Statistical Learning in Computer Vision, 2004.
12. Gray RM. Vector quantization. IEEE ASSP Magazine. 1984, 1(2):4-29.
13. Bay H, Ess A, Tuytelaars T, Van Gool L. SURF: speeded up robust features. Comput Vis Image Und. 2008, 110(3): 346-359.
14. Agresti A, Coull BA. Approximate is better than 'exact' for interval estimation of binomial proportions. Am Stat. 1998, 52:119–126.