BMC Bioinformatics

RESEARCH Open Access

Prediction of tumor location in prostate cancer tissue using a machine learning system on gene expression data

Osama Hamzeh[1†], Abedalrhman Alkhateeb[1*†], Julia Zheng[1], Srinath Kandalam[2] and Luis Rueda[1*†]

From 6th International Work-Conference on Bioinformatics and Biomedical Engineering
Granada, Spain. 25-27 April 2018

Abstract

Background: Finding the tumor location in the prostate is an essential pathological step for prostate cancer diagnosis and treatment. The location of the tumor – the laterality – can be unilateral (the tumor is affecting one side of the prostate), or bilateral on both sides. Nevertheless, the tumor can be overestimated or underestimated by standard screening methods. In this work, a combination of efficient machine learning methods for feature selection and classification are proposed to analyze gene activity and select them as relevant biomarkers for different laterality samples.

Results: A data set that consists of 450 samples was used in this study. The samples were divided into three laterality classes (left, right, bilateral). The aim of this work is to understand the genomic activity in each class and find relevant genes as indicators for each class with nearly 99% accuracy. The system identified groups of differentially expressed genes (RTN1, HLA-DMB, MRI1) that are able to differentiate samples among the three classes.

Conclusion: The proposed method was able to detect sets of genes that can identify different laterality classes. The resulting genes are found to be strongly correlated with disease progression. HLA-DMB and EIF4G2, which are detected in the set of genes can detect the left laterality, were reported earlier to be in the same pathway called Allograft rejection SuperPath.

Keywords: Machine learning, Classification, Biomarkers, Prostate cancer laterality

Background

Cancer is among the leading causes of death worldwide. In 2013, there were 8.2 million deaths, and 14.9 million cases of cancer incidence [1]. As with all cancer diseases, investigating prostate cancer at the molecular level reveals transcriptional and regulatory mechanisms of the tumour biology. Traditionally, prostate cancer studies centered primarily on finding biomarkers for differentiation between benign and cancerous tumors. Recently, studies have considered some other aspects of the tumours including progression, metastasis, location, and recurrence, among others.

Traditional methods for detecting prostate cancer such as prostate specific antigen (PSA) blood test, transrectal ultrasound image (TRUS) guided biopsy, and digital rectal exam (DRE) do not measure up to the medical standards. PSA blood test statistical results shows a specificity of 61% and a low sensitivity of 34.9%, while TRUS-guided biopsy and DRE are invasive [2].

In addition, multiparametric magnetic resonance imaging (MRI) of the prostate is a functional form of imaging used to augment standard T1- and T2-weighted imaging. Multiparametric MRI may miss up to 12% of cancer cases [3]. In addition to the need for reducing the number of

*Correspondence: alkhate@uwindsor.ca; Lrueda@uwindsor.ca
†Osama Hamzeh, Abedalrhman Alkhateeb, and Luis Rueda contributed equally to this work.
[1]School of Computer Science, University of Windsor, 401 Sunset Ave, N9B 3P4 Windsor, ON, Canada
Full list of author information is available at the end of the article

biopsies come most of the time with pain, fever, bleeding, infection, transient urinary difficulties, or other complications that require hospitalization [4]. Finding gene biomarkers of prostate cancer location and analyzing their proteomics can help clinically understand the development of the disease and improve treatment efficiency.

Machine learning approaches, on the other hand, have been successfully applied on prostate cancer data to identify gene biomarkers of the disease [5]. Using next generation sequencing and the power of machine learning, alkhateeb et al. devised a support vector machine (SVM) classifier to identify biomarker genes associated with prostate cancer progression. The biomarkers were able to discriminate consecutive prostate cancer stages with high performance [5]. Earlier, Hamzeh et al. proposed a method for finding groups of transcripts that are differentially expressed among the different Gleason stages [6]. The identified transcripts can be used to predict the actual Gleason score for new samples, and these transcripts belong to genes that are well known to play important roles in prostate and other types of cancer. Yu et al. demonstrated that their method is efficient for predicting prostate cancer aggressiveness based on gene expression patterns [7].

Similarly, machine learning approaches have been used for cancer localization prediction [8, 9]. Artan et al. proposed a prediction model based on a cost-sensitive SVM. The model is used to analyze a large data set of multispectral magnatic resonance imaging (MRI). This method improves the cost-sensitive SVM using a segmentation method by combining conditional random fields (CRF) with a cost-sensitive framework. Incorporating spatial information leads to better localization accuracy [8]. As stated earlier, prediction by imaging is still inaccurate, not specific and hence needs more improvement. In an attempt to find different gene expression levels between two lists, the first contains the expression levels of colon tumor cells, while the latter for rectal tumor cells, Sanz-Pamplona et al. applied agglomerative hierarchical clustering to display the classification ability between both lists. Both lists have very similar gene expression levels except for several HOX genes which are found to be associated with tumor location [9].

In this work, we are extending our previous method for classifying different laterality prostate samples which are left unary, right unary, or bilateral [10]. The results of this multi-class model are set of genes that can determine a specific class from the others. The literature shows that these genes are related to prostate cancer, which may lead to be a potential biomarkers for prostate cancer laterality.

Materials and methods

RNA-sequencing data from The Cancer Genome Atlas (TCGA) Prostate Adenocarcinoma (PRAD) was used. This data set consists of 450 samples for different patients with different cancer locations. There are three primary locations that the tumor might be located within the prostate: left, right and bilateral. Figure 1 shows the actual possible locations, while Table 1 describes the number of samples in each location.

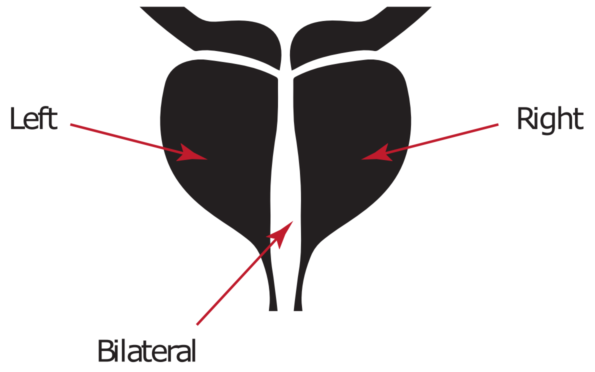

Fig. 1 Possible locations of the tumor in prostate cancer

Gene expression data was downloaded through the cBioPortal for cancer genomics database [11]. Each sample contains expression levels for each of the 60,488 genes; the gene expressions are given in terms of Transcripts Per Kilobase Million (TPM) values. The aim of this study is to identify genes which are associated with specific tumor locations, and hence we need to use the genes as features and the actual locations as classes to build a model to predict locations for future samples. Since most of the samples are bilateral, we deal with a class imbalance problem. We used the re-sampling method proposed in [12] as measure to lower the effect of this imbalance. The gene expressions of each node in the classification models are included in Additional file 1 that contains Bilateral-vs-Rest gene's expressions, Additional file 2 that contains Left-vs-Rest gene's expressions and Additional file 3 that contains Right-vs-Rest gene's expressions.

Re-sampling

By observing Table 1, we clearly notice that there is a class imbalance problem, where the number of samples in the right class (38) is almost twice as large as that of the left class (18). while the number of samples of the bilateral class (431) is more than twenty times larger than the left class and more than ten times larger than the right class.

Table 1 Number of samples in each prostate cancer tumor location

Left	Bilateral	Right
18	431	38

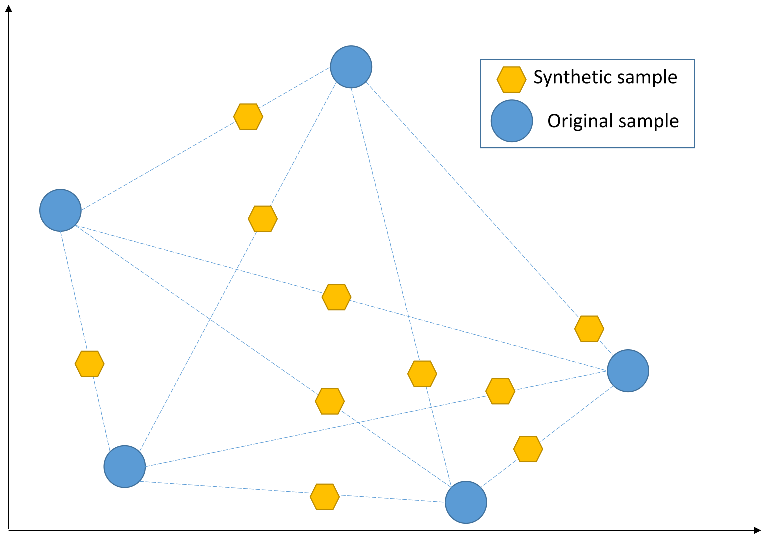

Fig. 2 Synthetic Minority Oversampling Technique (SMOTE) works by adding new synthetic sample randomly along the line that connects each of the two original samples

To solve this problem, multiple re-sampling methods were deployed and tested to identify a method that would yield the best solution for our data set. Oversampling provides a fast solution for classes left and right. This method duplicates samples from the minority classes and adds them until yielding a similar number of samples for each class. Applying oversampling directly did resolve the class imbalance problem and provided high accuracy for classifiers, although after taking a closer look at the samples used in these classifiers, we noticed that there was a major over-fitting. Based on the literature [13, 14], we selected the combination of oversampling Synthetic Minority Oversampling Technique (SMOTE) [15] and Neighborhood Cleaning Rule (NCL) [16] for under-sampling the majority class. Junsomboon et al. reported that the combination (NCL+SMOTE) outperformed another set of methods for handling the imbalance data sets. They have applied this combination on different health related data sets [13]. NCL uses the Wilson's Edited Nearest Neighbor Rule (ENN) to remove majority class outliers [17]. Batista et al. reported a high performance for SMOTE+ENN in handling imbalance data set [14].

NCL works by removing any sample whose class is different from the class of at least two of its three nearest neighbors. SMOTE introduces a new way of creating new samples, by utilizing the feature vector connecting each sample and introducing a new synthetic sample along the line that connects the two underlying samples. The exact location of the new sample on the line itself is calculated by measuring the distance between the two samples and multiplying that value by a random number between 0 and 1. Figure 2 shows the behavior of SMOTE.

Applying these two methods allowed us to use three classes that are balanced. Table 2 shows the number of samples after applying the SMOTE+ENN re-sampling methods.

Feature selection

Dealing with a huge number of features lead us to the problem of curse of dimensionality. As such, we use machine learning techniques to lower the number of features used for classification. We applied the information gain (IG) feature selection method [18] to rank all the genes with a score that relates to the highest information gain against the different classes. We then chose the attributes with the highest scores, discarding those with lower scores. In this paper, the IG attribute evaluator [18] is used to evaluate each attribute. IG of feature X with respect to class Y is calculated as follows:

$$IG(Y,X) = H(Y) - H(Y|X) \tag{1}$$

where,

$$H(Y) = -\sum_{y \in Y} p(y) log_2(p(y)). \tag{2}$$

and

$$H(Y|X) = -\sum_{x \in X} p(x) \sum_{y \in Y} p(y|x) log_2(p\,(y|x)). \tag{3}$$

Table 2 Number of samples in each prostate cancer tumor location after applying the SMOTE+ENN resampling methods

240	240	38	38	240	240
Left vs Bilateral		Left vs Right		Right vs Bilateral	

Table 3 Accuracy and precision for classifying each class versus the rest

Classifier	Accuracy	Precision	Accuracy	Precision	Accuracy	Precision
SVM RBF	99	97	99	97	99	97
Naive Bayes	88	78	82	78	80	78
Random Forest	93	85	90	85	95	85
	Left vs rest		Bilateral vs rest		Right vs rest	

Here, *H(Y)* is the entropy of class Y and *H(Y|X)* is the conditional entropy of Y given X.

The next step is to choose the best set of attributes (genes) that provide good classification among the different classes.

A wrapper that binds feature selection and classification methods is used. The feature selection method is the minimum redundancy maximum relevance (mRMR), which takes features that contain minimum redundancy while at the same time have high correlation to the classification variable [19]. The equation for minimizing redundancy (W_i) and maximizing the relevancy (V_i) is the following:

$$min\; W(S),\; W = \frac{1}{|S|^2} \sum_{i,j\in S} I(i,j), \tag{4}$$

and

$$max\; V(S,h),\; V = \frac{1}{|S|} \sum_{i\in S} I(h,i), \tag{5}$$

Where S is the set of features, *I(i,j)* is mutual information between features *(i,j)*, *h* is the class.

The operator ϕ(W,V) is defined to combine W and V and consider the following simplest form to optimize W and V at the same time:

$$max\, \phi(D,R), \phi = W - V, \tag{6}$$

Classification

We deal with a multi-class classification problem which is solved by using the one-versus-all approach. We have three different classes which are the three different locations. To apply the one-versus-all approach, we need to create three separate copies from the actual data set. For each data set, we set one of the classes to positive, and the rest of the classes are combined together to form the negative class. We used accuracy, sensitivity and specificity to choose the best classification method.

Multiple classification methods were applied on the data to identify which methods separate the locations better. Accordingly, the probabilistic classifier Naive Bayes that applies Bayes' theorem with the assumption of independence between the features [20] was tested. SVM was also used to build a classification model based on the features selected in the previous step [21]. The other classifier that was tested is random forest [22], which attempts to build multiple decision tree models with different samples and different initial variables.

The Weka open source tool were used to run different classification algorithms on the minimized number of features to identify which genes are differentially expressed in the different locations [23].

Results and discussion

The different classifiers produced varied results as observed in Table 3 and Fig. 3. The classifiers were chosen based on the accuracy and the precision, as leading high accuracy with low precision is not a good criterion at all. The accuracy measures the number of correctly classified samples divided by the number of all samples, while the precision is the true positive rate which measures the number of true positive calls divided by all positive calls. Table 3 shows the actual accuracy and precision for each classifier. The highest accuracy and precision for the different classifiers came from the SVM Radial basis function kernel (SVM-RBF) classifier. Grid search optimization was applied to fine tune the RBF classifier, it was able to separate the different locations by an accuracy of 99%. Random forest managed to result in high accuracy too, while the naive Bayes classifier results were not satisfactory.

Table 4 show the actual genes that were identified by SVM-RBF. These genes can be used to predict the location of the prostate cancer tumor very accurately from gene expression data.

Throughout our model 10-fold cross-validation was used. The proposed method identified 12 genes that are differentially expressed among the three different possible locations.

It is important to highlight that most of the genes identified in this work have been previously characterized and described to play some role in prostate cancer as well as other types of cancer. SNAI2 is a gene shown [24] to be silenced in prostate cancer and regulates neuroendocrine differentiation, metastasis-suppressor, and pluripotency gene expression.

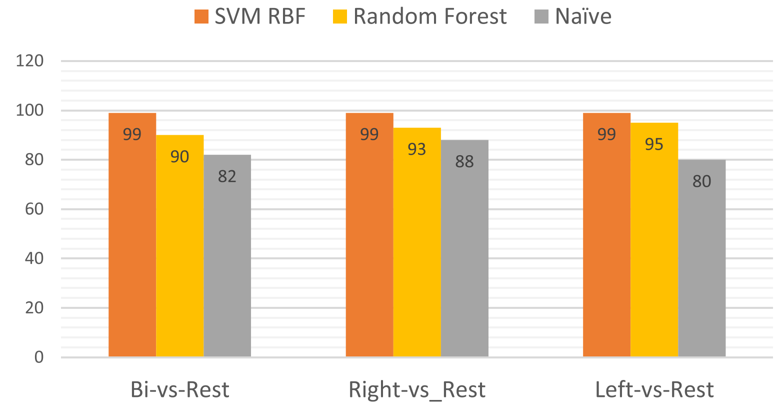

Fig. 3 Different classifiers accuracy for the different locations

Likewise, the results shown in [25, 26] indicate that increased TAF1/7 expression is associated with progression of human prostate cancers to the lethal castration-resistant state. In a similar way, the results reported in [27] found that tumor cell expression of HLA-DMB is associated with increased numbers of tumor-infiltrating CD8 T lymphocytes and both are associated with improved survival in advanced serous ovarian cancer.

Figures 4, 5, and 6 depict the ROC curves for all the classes versus the rest at each node. The area under the curve AUC for SVM-RBF tends to be further towards the north west with 0.99 value in the three figures, which means the best overall performance across all classes versus the rest. All other classifiers were inconsistent in the three figures. However, random forest performed very well in later false positive rates for both left and right classes with overall performance 0.87, 0.84 in order for both classes. it slightly outperformed the SVM-RBF in one point at both classes. but as we stated earlier, it was inconsistent through out different running parameters for false positive rates.

Biological insight

We have conducted a thoroughly literature review on the most up to date classification, as well as in the relevant databases and gathered valuable information about the most relevant genes that we have found in our study. A summary for each gene is given below and opens the avenue for future studies as well as additional lab experiments that can corroborate our studies and lead to novel ways of diagnosis, treatment and prognosis of the disease.

FBXO21 (F-box protein 21) is part of the multi-protein complex, SCF E3-ligase, which functions in phosphorylation-dependent ubiquitination. FBXO21 may affect prostate cancer through different mechanisms, and here we hypothesize two possibilities. Firstly, ABCB1 is a known tumour drug resistance biomarker because it is a multi-drug efflux pump linked with the development of metastases [28]. FBXO21 tags ABCB1 for proteasomal degradation, whereas inhibition of FBXO21 leads to higher expression level of ABCB1. Secondly, FBXO21 recognizes EID1 in cycling and G0 stage cells and targets it for degradation. EID1 interacts with retinoblastoma tumour suppressor (pRB), melanoma-associated antigen (MAGE), and E1A binding protein p300 (EP300) as well as being involved in the coupling cell cycle exit to cellular differentiation. All available evidence suggests that FBXO21 may be down-regulated in prostate cancer, although further research is desirable [29].

Table 4 Genes that can predict tumors in each location class of the prostate tumor

Ensemble	Gene	Ensemble	Gene	Ensemble	Gene
ENSG00000135108.13	FBXO21	ENSG00000120697.7	ALG5	ENSG00000242574.7	HLA-DMB
ENSG00000139970.15	RTN1	ENSG00000279453.1	Z99129	ENSG00000124193.13	SRSF6
ENSG00000128609.13	NDUFA5	ENSG00000019549.7	SNAI2	ENSG00000110321.14	EIF4G2
ENSG00000172336.4	POP7	ENSG00000037757.12	MRI1		
		ENSG00000178913.7	TAF7		
Left vs rest		Bilateral vs rest		Right vs rest	

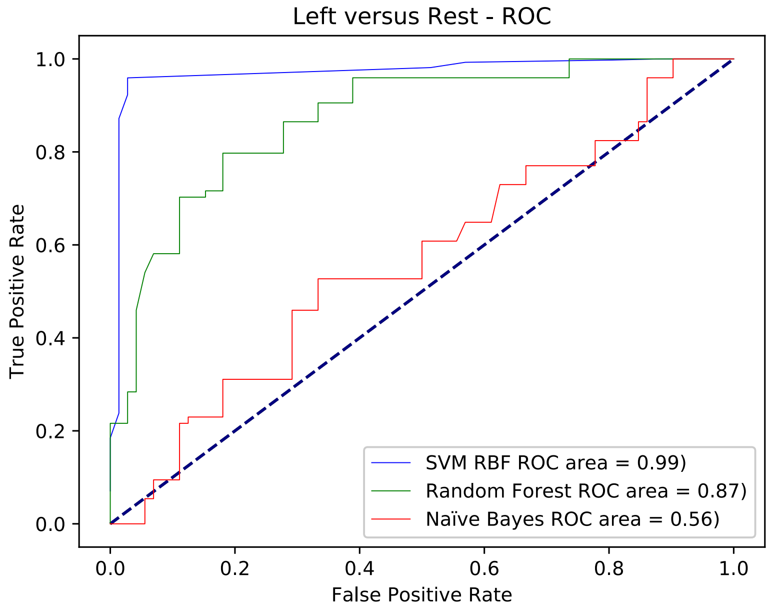

Fig. 4 The ROC curve for left versus the rest using different classifiers

RTN1 (reticulon 1) is associated with the endoplasmic reticulum (ER) and is involved in neuroendocrine secretions and membrane trafficking. RTN1 has been known exert a cancer-specific proapoptotic function. Specifically, RTN1-C regulates the two mutually exclusive ER stress-induced apoptosis and DNA damage-induced cell death. Over-expression of RTN1-C results in ER stress-induced cell death mediated by aberrantly increased cytosolic Ca^{2+} due to depletion of ER calcium stores [30]. A recent publication on prostate cancer shows that silencing RTN1 by siRNA enabled androgen-independent proliferation of androgen-dependent prostate cancer tumours. The knockdown of RTN1 increases the nuclear concentration of HDAC8, a multi-functional histone deacetylase that regulates activity of transcription factors such as nuclear hormone receptors [31]. In particular, it is known that ceramide inhibits androgen receptor activity and inhibits androgen-independent growth by

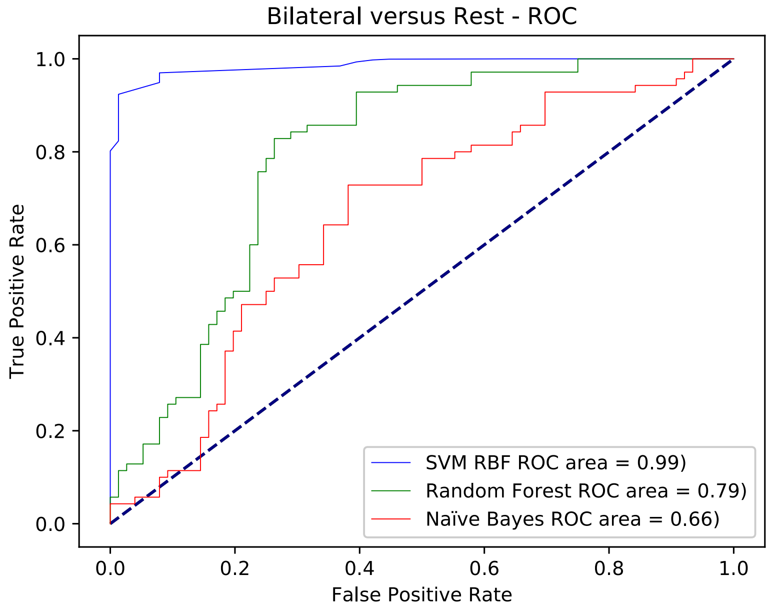

Fig. 5 The ROC curve for belateral versus the rest using different classifiers

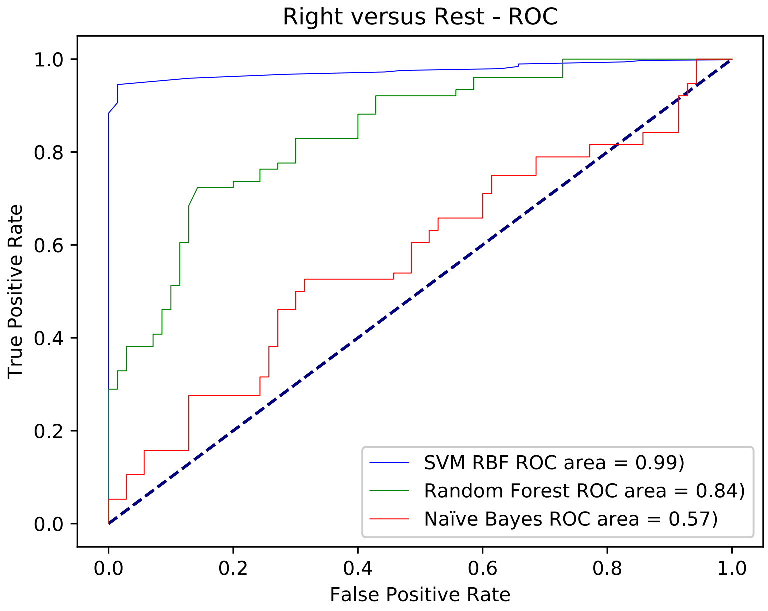

Fig. 6 The ROC curve for right versus the rest using different classifiers

activation of protein phosphatase 2A (PP2A) [32]. However, HDAC8-induced depletion of SPTSSA in the ER compromises the ER-localized ceramide biosynthesis pathway, leading to down-regulation of ceramide, partial inhibition of PP2A and androgen receptor activation in androgen-deprived conditions [31]. Consequently, RTN1 may be a proto-oncogene associated with aggressive, malignant and androgen-independent prostate cancer.

NDUFA5 (NADH:ubiquinone oxidoreductase subunit A5) is localized to the inner mitochondrial membrane and functions in the NADH two-electron reduction of ubiquinone [33]. Complex I, also known as NADH-ubiquinone oxidoreductase, is the first complex of the mitochondrial oxidative phosphorylation (OXPHOS) system. The energy released is coupled with generation of the electrochemical gradient necessary for ATP synthesis [34]. As expected, NDUFA5 activity is lower in hypoxic cells [35]. The Warburg effect states that tumour cells demonstrate drastically increased glycolysis activity compared to oxidative phosphorylation due to target genes up-regulated by hypoxia-inducible factor (HIF) [36]. On the other hand, NDUFA5 is up-regulated in HPV+ cervical cancer and its over-expression may play a role in carcinogenesis through acquiring growth advantage and resistance against an apoptotic signal [33]. In a recent publication, NDUFA5 also gained copy numbers in both low-grade and high-grade gliomas. Therefore, NDUFA5 may also be up-regulated in prostate cancer, although further research is necessary to confirm this hypothesis [37].

POP7 (POP7 homolog, ribonuclease P/MRP subunit) is discovered in S. cerevisiae. POP7 heterodimerizes to POP6 and binds to the P3 domain of catalytic ribonucleoproteins RNase MRP (mitochondrial RNA processing) and Rpr1 RNA [38]. RNase MRP is critically important to the viability of eukaryotic cells because it is localized in the nucleolus and is involved in processing mitochondrial RNAs and regulating mitochondrial DNA replication [39]. POP1/POP6/POP7 complex is required for telomere elongation protein (Est1) to associate with the RNP, which is critical during the process of mitosis for the cell lifespan before its senescence [40]. Despite the critical importance of POP7, no known human diseases are associated with this gene currently. Further research will be important to explore the biological significance of POP7.

HLA-DMB (major histocompatibility complex class II, DM beta) is a sub-unit of the HLA class II heterodimer found embedded in intracellular vesicles. In antigen-presenting cells (APC), HLA-DMB is critical in the antigen-presentation machinery by releasing class II-associated invariant chain peptide (CLIP) from MHC class II molecules so that the peptide binding site is free to interact with antigenic peptides [41]. A recent publication on prostate cancer research found that HLA-DMB is co-expressed with ERG and silencing ERG led to significant under-expression of HLA-DMB. Thus, HLA-DMB is an up-regulated tumour-associated gene in prostate cancer [42].

SRSF6 (Serine and Arginine rich Splicing Factor 6) modulates a splicing factor protein called SFRS12 to determine alternative splicing of mRNA. In a recent publication on colorectal cancer, SRSF6 targeted ZO-1 (tight junction protein 1) exon23 for alternative splicing, consequentially disrupting ZO-1 from regulating tight junctions between adjacent cells [43]. Furthermore, SRSF6 is the direct target of LINC01133, a key SRSF6 modulates a splicing factor protein called SFRS12 to determine alternative splicing of mRNA. In a 2017 paper on colorectal cancer, SRSF6 targeted ZO-1 (tight junction protein 1) exon23 for alternative splicing, consequentially disrupting ZO-1 from regulating tight junctions between adjacent cells. In addition, SRSF6 is the direct target of LINC01133, a key downstream protein of TGF-β signaling pathway which is critical for cell growth and differentiation [44]. Silencing SRSF6 in colorectal cancer tissues inhibited epithelial-mesenchymal transition, tissue invasion, and metastasis. A study on wound healing found that over-expression of SRSF6 induces skin hyperplasia due to SRSF6 up-regulating Tenascin C and suppressing the normal epithelial differentiation mechanism. Therefore, SRSF6 may be up-regulated in prostate cancer [43].

EIF4G2 gene, Eukaryotic Translation Initiation Factor 4 Gamma 2 is a cap - binding protein complex which has three sub units – eiF4A, eiF4E eiF4G. The gene is known to up-regulate p21, a cyclin dependant kinase inhibitor and interleukin 6 [45]. Higher expression levels of p21 oncogene protein are found with increasing prostate cancer tumor grade [46]. Interleukin 6 is involved in the progression of prostate cancer [47], and is used as a clinicopathological feature by detecting the levels in serum [48]. With the up-regulated expression levels of EIF4G2 gene in prostate cancer, it can be used as a potential marker for studying the progression of the disease.

Interestingly, EIF4G2 and HLA-DMB which are part of the gene set that can identify right side from the rest, they are both part of Allograft rejection SuperPath pathway [49].

The discovery of fusion protein transcripts in the recent times have helped studying prostate cancer development with much detail. ALG5, Dolichyl-Phosphate Beta-Glucosyltransferase and PIGU, Phosphatidylinositol Glycan Anchor Biosynthesis Class forms a chimeric-fusion protein transcript in which glucosyltransferase, the head from ALG5 is retained but GPI transamidase, the tail has been eliminated in PIGU resulting in the loss of functionality of both the genes [50]. The uncommon joining of the genes would result in serious complications in the overall environment of the cell causing further progression of the cancer. The transcription of the fused ALG5-PIGU is androgen independent [51]. Fusion protein transcripts will serve as an important biomarker both in detection and treatment of Prostate Cancer.

SNAI2, Snail Family Transcriptional Repressor 2 encodes zinc-finger protein of the Snail family transcription factors, is involved in the generation and migration of neural crest cells in embryonic stages which is driven by epithelial to mesenchymal transition (EMT). Presence of neuroendocrine cells in nests - neuroendocrine differentiation (NED) is a known histological marker for prostate Cancer. SNAI2 expression is down regulated in prostate cancer and silencing of the gene may turn on neuroendocrine differentiation, pluripotent genes and turn on specific metastasis suppressors [52]. SNAI2 knockdown initiating metastatic suppressor genes involves many pathways and further research is needed to derive a conclusion. Studies of SNAI2 gene regulation properties will help us in understanding the development of prostate cancer.

MRI1, Methylthioribose-1-Phosphate Isomerase 1 gene helps in catalyses of methionine, an important amino acid, in methionine salvage pathway. Development of certain cancers like prostate, glioma, bladder, breast, melanoma are dependent on methionine [53, 54]. To understand the dependency of methionine in prostate cancer a study has been conducted on patients who were not receiving any conventional treatment and were undergoing an intensive lifestyle program with a restricted methionine vegan diet. Analysis of serum samples revealed that there was a 70% inhibition of the growth androgen sensitive prostate adenocarcinoma (LNCaP) cells [55]. The data suggests that methionine restricted diet and lifestyle changes may help in slowing down the development of prostate cancer.

Conclusion

Understanding gene activity in the prostate cancer laterality my help to guide the diagnosis and treatment of the disease. In this work, we have proposed a machine learning method that is capable of predicting with a high accuracy the tumor location in a cancer infected prostate. As a result, we have found genes as indicators that can differentiate the three locations of prostate cancer with high accuracy. The contributions of this study are two-fold. The proposed machine learning system can be used as a protocol for other types of cancer and other clinical problems in cancer studies. It also open the doors for potential biomarkers that can be further tested in wet-lab scenarios with the hope to move to clinical trials in order to replace the invasive biopsy or inaccurate image scanning.

The literature shows strong relations between prostate cancer metastasis and the computationally derived genes. Wet-lab experiments and RNA-seq profiling of those genes will better explore the relation between the findings and the prostate cancer laterality, which will potentially help the prognosis of the disease.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s12859-020-3345-9.

Additional file 1: Bilateral-vs-Rest gene's expressions.

Additional file 2: Left-vs-Rest gene's expressions.

Additional file 3: Right-vs-Rest gene's expressions.

Abbreviations

Genes that can predict tumors in each location class of the prostate tumor. AUC: Area Under the Curve; DRE: Digital Rectal Exam; CRF: Conditional Random Field; ENN: Edited Nearest Neighbor; ER: Endoplasmic Reticulum; FBXO21: F-BOX Protein 21; HIF: Hypoxia-Inducible Fact; IG: Information Gain; MAGE: Melanoma-Associated Antige; MRI: Multiparametric Magnetic Resonance Imaging; mRMR: minimum Redundancy Maximum Relevance; MRP: Mitochondrial RNA Processing; NCL: Neighborhood Cleaning; NDUFA5: NADH:Ubiquinone Oxidoreductase Subunit A5; PP2A: Protein Phosphatase 2A; PRAD: Prostate Adenocarcinoma; PSA: Prostate Specific Antigen; RBF: Radial Basis Function; ROC: Receiver Operating Characteristic curve; SMOTE: Synthetic Minority Oversampling Technique; SRSF6: Serine and Arginine Rich Splicing Factor 6; SVM: Support Vector Machine; TCGA: The Cancer Genome Atlas; TPM: Transcripts Per kilobase Million; TRUS: Transrectal Ultrasound

Acknowledgements

This research has been partially supported by NSERC, the Natural Sciences and Engineering Research Council of Canada. The authors would like to thank the School of Computer Science and the Office of Research Services at the University of Windsor for their financial support too. This research was enabled in part by support provided by SHARCNET (http://www.sharcnet.ca) and Compute Canada (http://www.computecanada.ca).

About this supplement

This article has been published as part of *BMC Bioinformatics Volume 21 Supplement 2, 2020: Selected articles from the 6th International Work-Conference on Bioinformatics and Biomedical Engineering*. The full contents of the supplement are available online at URL.

Authors' contributions

OH, AA participated equally in implementing the methods, they discussed the idea and verified the model with LR. JZ and SK verified the biological findings, all authors have participated in writing the paper. LR is a principal investigator who led the main project. All authors read and approved the final manuscript.

Funding

Publication costs are funded by the Natural Sciences and Engineering Research Council of Canada, NSERC.

Availability of data and materials

The datasets analyzed during the current study are available in the cBioPortal repository at http://www.cbioportal.org/study?id=prad_tcga.

Ethics approval and consent to participate

Not applicable.

Consent for publication

All authors have read and approved of the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]School of Computer Science, University of Windsor, 401 Sunset Ave, N9B 3P4 Windsor, ON, Canada. [2]Department of Biomedical Sciences, University of Windsor, 401 Sunset Ave, N9B 3P4 Windsor, ON, Canada.

Published: 13 March 2020

References

1. Stewart B, Wild P, et al. World cancer report 2014. Health. 2017. http://publications.iarc.fr/Non-Series-Publications/World-Cancer-Reports/World-Cancer-Report-2014.
2. Parpart S, Rudis A, Schreck A, Dewan N, Warren P. Sensitivity and specificity in prostate cancer screening methods and strategies. J Young Investig. 2007. http://www.jyi.org/issue/sensitivity-and-specificity-in-prostate-cancerscreening-methods-and-strategies/.
3. Stewart W, Lizama S, Peairs K, Sateia F, Choi Y. Screening for prostate cancer. In: Seminars in Oncology. Elsevier; 2017.
4. Rosario J, Lane J, Metcalfe C, Donovan L, Doble A, Goodwin L, Davis M, Catto W, Avery K, Neal E, et al. Short term outcomes of prostate biopsy in men tested for cancer by prostate specific antigen: prospective evaluation within protect study. Bmj. 2012;344:d7894.
5. Alkhateeb A, Rezaeian I, Singireddy S, Cavallo-Medved D, Porter LA, Rueda L. Transcriptomics signature from next-generation sequencing data reveals new transcriptomic biomarkers related to prostate cancer. Cancer Informat. 2019;18:1176935119835522.
6. Hamzeh O, Alkhateeb A, Rueda L. Finding Transcripts Associated with Prostate Cancer Gleason Stages Using Next Generation Sequencing and Machine Learning Techniques. In: International Conference on Bioinformatics and Biomedical Engineering. Cham: Springer; 2017.
7. Ping Y, Landsittel D, Jing L, Nelson J, Ren B, Liu L, McDonald C. Gene expression alterations in prostate cancer predicting tumor aggression and preceding development of malignancy. J Clin Oncol. 2004;22(14):2790–9.
8. Artan Y, Haider A, Langer L, Kwast H, Evans J, Yang Y, Wernick N, Trachtenberg J, Yetik I. Prostate cancer localization with multispectral mri using cost-sensitive support vector machines and conditional random fields. IEEE Trans Image Process. 2010;19(9):2444–5.
9. Sanz-Pamplona R, Cordero D, Berenguer A, Lejbkowicz F, Rennert H, Salazar R, Biondo S, Sanjuan X, Pujana A, Rozek L. Gene expression differences between colon and rectum tumors. Clin Cancer Res. 2011;17(23):7303–12.
10. Hamzeh O, Alkhateeb A, Rueda L. Predicting Tumor Locations in Prostate Cancer Tissue Using Gene Expression. In: Rojas I, Ortuño F, editors. Bioinformatics and Biomedical Engineering. IWBBIO 2018. Lecture Notes in Computer Science, vol 10813. Cham: Springer; 2018.
11. GDC. Portal.gdc.cancer.gov. 2017. https://portal.gdc.cancer.gov/. Accessed 15 Aug 2017.
12. Estabrooks A, Jo T, Japkowicz N. A multiple resampling method for learning from imbalanced data sets. Comput Intell. 2004;20(1):18–36.
13. Junsomboon N, Phienthrakul T. Combining Over-Sampling and Under-Sampling Techniques for Imbalance Dataset. In: Proceedings of the 9th International Conference on Machine Learning and Computing. New York: Association for Computing Machinery; 2017. p. 243–7. https://doi.org/10.1145/3055635.3056643.
14. Batista G, Prati R, Monard M. A study of the behavior of several methods for balancing machine learning training data. ACM SIGKDD Explor Newsl. 2004;6(1):20–9.
15. Chawla N, Bowyer K, Hall O, Kegelmeyer P. SMOTE: synthetic minority over-sampling technique. J Artif Intell Res. 2002;16:321–57.
16. Laurikkala J. Improving Identification of Difficult Small Classes by Balancing Class Distribution. Tech. Rep. A-2001-2, University of Tampere. 2001.
17. Wilson D. Asymptotic properties of nearest neighbor rules using edited data. IEEE Trans Syst Man Cybern. 1972;SMC-2(3):408–21.
18. Novakovic J. Using information gain attribute evaluation to classify sonar targets. In: 17th Telecommunications forum TELFOR. Belgrade; 2009. p. 24–6.
19. Peng H, Long F, Ding C. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans Patt Anal Mach Intell. 2005;27(8):1226–38.
20. Domingos P, Pazzani M. On the optimality of the simple Bayesian classifier under zero-one loss. Mach Learn. 1997;29(2-3):103–30.
21. Cortes C, Vapnik V. Support-vector networks. Mach Learn. 1995;20(3): 273–97.
22. Rodriguez F, Ghimire B, Rogan J, Olmo M, Sanchez P. An assessment of the effectiveness of a random forest classifier for land-cover classification. ISPRS J Photogramm Remote Sens. 2012;67:93–104.
23. Frank E, Hall M, Witten I. The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Tech-niques" 4th ed: Morgan Kaufmann; 2016.

24. Esposito S, Russo V, Airoldi I, Tupone G, Sorrentino C, Barbarito G, Di Carlo E. SNAI2/Slug gene is silenced in prostate cancer and regulates neuroendocrine differentiation, metastasis-suppressor and pluripotency gene expression. Oncotarget. 2015;6(19):1–34.
25. Tavassoli P, Wafa L, Cheng H, Zoubeidi A, Fazli L, Gleave M, Snoek R, Rennie P. TAF1 Differentially Enhances Androgen Receptor Transcriptional Activity via Its N-Terminal Kinase and Ubiquitin-Activating and -Conjugating Domains. Mol Endocrinol. 2010;24(4):696–708. https://doi.org/10.1210/me.2009-0229.
26. Bhattacharya S, Lou X, Hwang P, Rajashankar K, Wang X, Gustafsson J, Fletterick R, Jacobson R, Webb P. Structural and functional insight into TAF1–TAF7, a subcomplex of transcription factor II D. PNAS. 2014;111(25): 9103–8. https://doi.org/10.1073/pnas.1408293111.
27. Callahan M, Nagymanyoki Z, Bonome T, et al. Increased Hla-Dmb Expression In The Tumor Epithelium Is Associated With Increased Cytotoxic T Lymphocyte Infiltration And Improved Prognosis In Advanced Serous Ovarian Cancer. Clin Cancer Res. 2008;14(23):7667–73. https://doi.org/10.1158/1078-0432.CCR-08-0479.
28. Ravindranath A, Kaur S, Wernyj R, Kumaran M, Gonzalez K, Chan R, Lim E, Madura K, Rodriguez L. Cd44 promotes multi-drug resistance by protecting p-glycoprotein from fbxo21-mediated ubiquitination. Oncotarget. 2015;6(28):26308.
29. Zhang C, Li X, Adelmant G, Dobbins J, Geisen C, Oser M, Wucherpfenning K, Marto J, Kaelin W. Peptidic degron in eid1 is recognized by an scf e3 ligase complex containing the orphan f-box protein fbxo21. Proc Natl Acad Sci. 2015;112(50):15372–7.
30. Chen L, Wan L, Du J, Shen Y. Identification of manf as a protein interacting with rtn1-c. Acta Biochim Biophys Sin. 2014;47(2):91–7.
31. Levina E, Ji H, Chen M, Baig M, Oliver D, Ohouo P, Lim C, Schools G, Carmack S, Ding Y, et al. Identification of novel genes that regulate androgen receptor signaling and growth of androgen-deprived prostate cancer cells. Oncotarget. 2015;6(15):13088.
32. Bhardwaj A, Singh S, Srivastava S, Honkanen RE, Reed E, Singh AP. Modulation of protein phosphatase 2A activity alters androgenindependent growth of prostate cancer cells: therapeutic implications. Mol Cancer Ther. 2011;10(5):720–31.
33. Denninger K, Litman T, Marstrand T, Moller K, Svensson L, Labuda T, Andersson A. Kinetics of gene expression and bone remodelling in the clinical phase of collagen-induced arthritis. Arthritis Res Ther. 2015;17(1):43.
34. Sato Y, Inoue M, Yoshizawa T, Yamagata K. Moderate hypoxia induces β-cell dysfunction with hif-1–independent gene expression changes. PLoS ONE. 2014;9(12):e114868.
35. Peralta S, Torraco A, Wenz T, Garcia S, Diaz F, Moraes C. Partial complex i deficiency due to the cns conditional ablation of ndufa5 results in a mild chronic encephalopathy but no increase in oxidative damage. Hum Mol Genet. 2013;23(6):1399–412.
36. Hsu P, Sabatini D. Cancer cell metabolism: Warburg and beyond. Cell. 2008;134(5):703–707.
37. Li Y, Wang D, Wang L, Yu J, Du D, Chen Y, Gao P, Wang D-M, Yu J, Zhang F, et al. Distinct genomic aberrations between low-grade and high-grade gliomas of chinese patients. PLoS ONE. 2013;8(2):e57168.
38. Reiner R, Alfiya-Mor N, Demma M, Wesolowski D, Altman S, Jarrous N. RNA binding properties of conserved protein subunits of human rnase p. Nucleic Acids Res. 2011;39(13):5704–14.
39. Esakova O, Krasilnikov A. Of proteins and rna: the rnase p/mrp family. RNA. 2010;16(9):1725–47.
40. Collins K. The biogenesis and regulation of telomerase holoenzymes. Nat Rev Mol Cell Biol. 2006;7(7):484.
41. Theodoridis S, Pikrakis A, Koutroumbas K, Cavouras D. Introduction to Pattern Recognition: A Matlab Approach: A Matlab Approach. Waltham: Academic Press; 2010.
42. Paulo P, Ribeiro F, Santos J, Mesquita D, Almeida M, Silva J, Itkonen H, Henrique R, Jerónimo C, Sveen A, et al. Molecular subtyping of primary prostate cancer reveals specific and shared target genes of different ets rearrangements. Neoplasia (New York, NY). 2012;14(7):600.
43. Wan L, Yu W, Shen E, Sun W, Liu Y, Kong J, Wu Y, Han F, Zhang L, Yu T, et al. SRSF6-regulated alternative splicing that promotes tumour progression offers a therapy target for colorectal cancer. Gut. 2019;68(1): 118–29.
44. Kim H, Lee G, Choi K, Kim D, Ryu J, Hwang K, Na K, Choi C, Hong Kuh J, Chung M, et al. Srsf5: a novel marker for small-cell lung cancer and pleural metastatic cancer. Lung Cancer. 2016;99:57–65.
45. Mori R, Xiong S, Wang Q, Tarabolous C, Shimada H, Panteris E, Danenberg K, Danenberg P, Pinski J. Gene profiling and pathway analysis of neuroendocrine transdifferentiated prostate cancer cells. The Prostate. 2009;69(1):12–23.
46. Viola M, Fromowitz F, Oravez S, Deb S, Finkel G, Lundy J, Hand P, Thor A, Schlom J. Expression of RAS oncogene P21 in prostate cancer. N Engl J Med. 1986;314(3):133–7.
47. Chung T, Yu J, Spiotto M, Bartkowski M, Simons J. Characterization of the role of il-6 in the progression of prostate cancer. The Prostate. 1999;38(3):199–207.
48. Michalaki V, Syrigos K, Charles P, Waxman J. Serum levels of il-6 and tnf-α correlate with clinicopathological features and patient survival in patients with prostate cancer. Br J Cancer. 2004;90(12):2312.
49. Belinky F, Nativ N, Stelzer G, Zimmerman S, Stein T, Safran M, Lancet D. Pathcards: multi-source consolidation of human biological pathways. Database. 2015;2015.
50. Luo J, Liu S, Zuo Z, Chen R, Tseng G, Yan P. Discovery and classification of fusion transcripts in prostate cancer and normal prostate tissue. Am J Pathol. 2015;185(7):1834–45.
51. Pflueger D, Terry S, Sboner A, Habegger L, Esgueva R, Lin P-C, Svensson MA, Kitabayashi N, Moss BJ, MacDonald TY, et al. Discovery of non-ETS gene fusions in human prostate cancer using next-generation RNA sequencing. Genome Res. 2011;21(1):56–67.
52. Esposito S, Russo M, Airoldi I, Tupone M, Sorrentino C, Barbarito G, Di Meo S, Di Carlo E. SNAI2/sLUG gene is silenced in prostate cancer and regulates neuroendocrine differentiation, metastasis-suppressor and pluripotency gene expression. Oncotarget. 2015;6(19):17121.
53. Cavuoto P, Fenech M. A review of methionine dependency and the role of methionine restriction in cancer growth control and life-span extension. Cancer Treat Rev. 2012;38(6):726–36.
54. Breillout F, Antoine E, Poupon M. Methionine dependency of malignant tumors: a possible approach for therapy. JNCI J Natl Cancer Inst. 1990;82(20):1628–32.
55. Ornish D, Weidner G, Fair W, Marlin R, Pettengill E, Raisin C, Dunn-Emke S, Crutchfield L, Jacobs F, Barnard R, et al. Intensive lifestyle changes may affect the progression of prostate cancer. J Urol. 2005;174(3):1065–70.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.