RESEARCH NOTE

Predicting Outcomes of Hormone and Chemotherapy in the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) Study by Biochemically-inspired Machine Learning [version 1; peer review: 2 approved with reservations]

Iman Rezaeian¹, Eliseos J. Mucaki², Katherina Baranova², Huy Q. Pham¹, Dimo Angelov³, Alioune Ngom¹, Luis Rueda¹, Peter K. Rogan^{id} 2-4

¹Department of Computer Science, University of Windsor, Windsor, Canada

²Department of Biochemistry, University of Western Ontario, London, Canada

³Department of Computer Science, University of Western Ontario, London, Canada

⁴CytoGnomix Inc, London, Canada

V1 First published: 31 Aug 2016, 5:2124
<https://doi.org/10.12688/f1000research.9417.1>

Second version: 27 Jan 2017, 5:2124
<https://doi.org/10.12688/f1000research.9417.2>

Latest published: 12 May 2017, 5:2124
<https://doi.org/10.12688/f1000research.9417.3>

Abstract

Genomic aberrations and gene expression-defined subtypes in the large METABRIC patient cohort have been used to stratify and predict survival. The present study used normalized gene expression signatures of paclitaxel drug response to predict outcome for different survival times in METABRIC patients receiving hormone (HT) and, in some cases, chemotherapy (CT) agents. This machine learning method, which distinguishes sensitivity vs. resistance in breast cancer cell lines and validates predictions in patients, was also used to derive gene signatures of other HT (tamoxifen) and CT agents (methotrexate, epirubicin, doxorubicin, and 5-fluorouracil) used in METABRIC. Paclitaxel gene signatures exhibited the best performance, however the other agents also predicted survival with acceptable accuracies. A support vector machine (SVM) model of paclitaxel response containing the *ABCB1*, *ABCB11*, *ABCC1*, *ABCC10*, *BAD*, *BBC3*, *BCL2*, *BCL2L1*, *BMF*, *CYP2C8*, *CYP3A4*, *MAP2*, *MAP4*, *MAPT*, *NR1I2*, *SLCO1B3*, *TUBB1*, *TUBB4A*, *TUBB4B* genes was 78.6% accurate in 84 patients treated with both HT and CT (median survival ≥ 4.4 yr). Accuracy was lower (73.4%) in 304 untreated patients. The performance of other machine learning approaches were also evaluated at different survival thresholds. Minimum redundancy maximum relevance feature selection of a paclitaxel-based SVM

Open Peer Review

Approval Status

	1	2
version 3		
(revision)
12 May 2017	view	
version 2		
(revision)
27 Jan 2017	view	view
version 1		
31 Aug 2016	view	view

1. **Elana J. Fertig**, Johns Hopkins University, Baltimore, USA

2. **Chun-Wei Tung** , Kaohsiung Medical University, Kaohsiung, Taiwan

Any reports and responses or comments on the article can be found at the end of the article.

classifier based on expression of *ABCB11*, *ABCC1*, *BAD*, *BBC3* and *BCL2L1* was 79% accurate in 53 CT patients. A random forest (RF) classifier produced a gene signature (*ABCB11*, *ABCC1*, *BAD*, *BCL2*, *CYP2C8*, *CYP3A4*, *MAP4*, *MAPT*, *NR1I2*, *TUBB1*, *GBP1*, *OPRK1*) that predicted >3 year survival with 82.4% accuracy in 420 HT patients. A similar RF gene signature showed 79.6% accuracy in 504 patients treated with CT and/or HT. These results suggest that tumor gene expression signatures refined by machine learning techniques can be useful for predicting survival after drug therapies.

Keywords

Gene expression signatures , breast cancer , chemotherapy resistance , hormone therapy , machine learning , support vector machine , random forest

This article is included in the **Bioinformatics** gateway.

This article is included in the **Machine learning: life sciences** collection.

Corresponding author: Peter K. Rogan (progan@uwo.ca)

Competing interests: PKR cofounded Cytognomix. A patent application related to biologically inspired gene signatures is pending. The other authors declare that they have no competing interests.

Grant information: AN and LR are funded by NSERC grants RGPIN-2016-05017 and RGPIN-2014-05084 and by the Windsor Essex County Cancer Centre Foundation under a Seeds4Hope grant. PKR has been supported by NSERC [Discovery Grant 371758-2009], Canadian Foundation for Innovation, Canada Research Chairs and Cytognomix Inc.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Copyright: © 2016 Rezaeian I *et al.* This is an open access article distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. Data associated with the article are available under the terms of the [Creative Commons Zero "No rights reserved" data waiver](#) (CC0 1.0 Public domain dedication).

How to cite this article: Rezaeian I, Mucaki EJ, Baranova K *et al.* **Predicting Outcomes of Hormone and Chemotherapy in the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) Study by Biochemically-inspired Machine Learning [version 1; peer review: 2 approved with reservations]** F1000Research 2016, 5:2124 <https://doi.org/10.12688/f1000research.9417.1>

First published: 31 Aug 2016, 5:2124 <https://doi.org/10.12688/f1000research.9417.1>

RF learning: RF was trained using the WEKA 3.7⁸ data mining tool. This classifier uses multiple random trees for classification, which are combined via a voting scheme to make a decision on the given input gene set. **Figure 2** depicts the therapy outcome prediction process of a given patient using a RF consisting of a series of decision trees derived from different subsets of paclitaxel-related genes.

Augmented Gene Selection: The most relevant genes (features) for therapy outcome prediction were found using the minimum redundancy and maximum relevance (mRMR) approach⁹. mRMR is a wrapper that incrementally selects genes by maximizing the average mutual information between gene expression features and classes, while minimizing their redundancies:

$$mRMR = \max_s \left[\frac{1}{|s|} \sum_{f_i \in s} I(f_i, C) - \frac{1}{|s|^2} \sum_{f_i, f_j \in s} I(f_i, f_j) \right]$$

where f_i corresponds to a feature in gene set S , $I(f_i, C)$ is the mutual information between f_i and class C , and $I(f_i, f_j)$ is the mutual information between features f_i and f_j .

For this experiment, we used a 26-gene signature (genes *ABCB1*, *ABCB11*, *ABCC1*, *ABCC10*, *BAD*, *BBC3*, *BCL2*, *BCL2L1*, *BMF*, *CYP2C8*, *CYP3A4*, *MAP2*, *MAP4*, *MAPT*, *NR1H2*, *SLCO1B3*, *TUBB1*, *TUBB4A*, *TUBB4B*, *FGF2*, *FN1*, *GBP1*, *NFKB2*, *OPRK1*, *TLR6*, *TWIST1*) as the base feature set. These genes were selected (in Ref. 1) based either on their known involvement in paclitaxel metabolism, or evidence that their expression levels and/or copy numbers correlate with paclitaxel GI_{50} values (Table 3). mRMR and

SVM were combined to obtain a subset of genes that can accurately predict patient survival outcome; here, we considered 3, 4 and 5 years as survival thresholds for breast cancer patients (Table 3).

Results and discussion

Dataset 1. Predicted treatment response for each individual METABRIC patient

<http://dx.doi.org/10.5256/f1000research.9417.d133983>

The predicted and expected response to treatment for each individual METABRIC patient for each analyses listed in Table 1, Table 2 and Table 3 are indexed. Patients sensitive to treatment are labeled with '0' while resistant patients are labeled '1'.

The performance of several ML techniques have been compared that distinguish paclitaxel sensitivity and resistance in METABRIC patients using its tumour gene expression datasets. SVMs have generated gene signatures, indicating which genes are important for treatment response in METABRIC patients. These models are more accurate for prediction of outcomes in patients receiving HT and/or CT compared to other patient groups.

SVMs and RF were trained using expression of genes associated with paclitaxel response, mechanism of action and stable genes in the biological pathways of these targets (Figure 3). SVM models for drugs used to treat these patients were derived by backwards feature selection on patient subsets stratified by treatment or outcome (Table 1). The highest SVM accuracy was found for the

Figure 2. RF decision tree diagram depicts the therapy outcome prediction process of a given patient, using a RF consisting of k decision trees. Several DTs are built using different subsets of paclitaxel-related genes. The process starts from the root of each tree and if the expression of the gene corresponding to that node is greater than a specific value, the process continues through the right branch, otherwise it continues through the left branch until it reaches a leaf node; that leaf represents the prediction of the tree for that specific input. The decisions of all trees are considered and the one with the largest number of votes is selected as the patient outcome.

Figure 3. Schematic elements of gene expression changes associated with response to paclitaxel. Red boxes indicate genes with a positive correlation between gene expression or copy number, and resistance using multiple factor analysis. Blue demonstrates a negative correlation. Genes outlined in dark grey are those in a previously published paclitaxel SVM model (reproduced from reference ¹ with permission).

Table 1. SVM gene expression signature performance on METABRIC patients.

Patient treatment	# of patients	Agent: final gene signature	Accuracy (%)
CT and CT/HT combination (without radiation therapy) ¹	84	Paclitaxel: <i>ABCC1, ABCC10, BAD, BIRC5, FN1, GBP1, MAPT, SLCO1B3, TMEM243, TUBB3, TUBB4B</i>	78.6
		Tamoxifen: <i>ABCC2, ALB, CCNA2, E2F7, FLAD1, FMO1, NCOA2, NR1I2, PIAS4, SULT1E1</i>	76.2
		Methotrexate: <i>ABCC2, ABCG2, CDK2, DHFRL1</i>	71.3
		Epirubicin: <i>ABCB1, CDA, CYP1B1, ERBB3, ERCC1, MTHFR, PON1, SEMA4D, TFDP2</i>	72.6
		Doxorubicin: <i>ABCC2, ABCD3, CBR1, FTH1, GPX1, NCF4, RAC2, TXNRD1</i>	75.0
		5-Fluorouracil: <i>ABCB1, ABCC3, MTHFR, TP53</i>	71.4
CT and/or HT ^{1,2,3,4}	735	Paclitaxel: <i>BAD, BCAP29, BCL2, BMF, CNGA3, CYP2C8, CYP3A4, FGF2, FN1, NFKB2, NR1I2, OPRK1, SLCO1B3, TLR6, TUBB1, TUBB3, TUBB4A, TUBB4B, TWIST1</i>	66.1
Deceased only ^{2,4,5} (CT and/or HT)	327	Paclitaxel: <i>ABCB11, BAD, BBC3, BCL2, BCL2L1, BIRC5, CYP2C8, FGF2, FN1, GBP1, MAPT, NFKB2, OPRK1, SLCO1B3, TMEM243</i>	75.2
No treatment ¹	304	Paclitaxel: <i>ABCB1, ABCB11, BBC3, BCL2L1, BMF, CYP3A4, FGF2, GBP1, MAP4, MAPT, NR1I2, OPRK1, SLCO1B3, TUBB4A, TUBB4B, TWIST2</i>	73.4

Initial gene sets preceding feature selection: Paclitaxel - *ABCB1, ABCB11, ABCC1, ABCC10, BAD, BBC3, BCAP29, BCL2, BCL2L1, BIRC5, BMF, CNGA3, CYP2C8, CYP3A4, FGF2, FN1, GBP1, MAP2, MAP4, MAPT, NFKB2, NR1I2, OPRK1, SLCO1B3, TLR6, TUBB1, TWIST1*. Tamoxifen - *ABCB1, ABCC2, ALB, C10ORF11, CCNA2, CYP3A4, E2F7, F5, FLAD1, FMO1, IGF1, IGF1BP3, IRS2, NCOA2, NR1H4, NR1I2, PIAS4, PPARA, PROC, RXRA, SMARCD3, SULT1B1, SULT1E1, SULT2A1*. Methotrexate - *ABCB1, ABCC2, ABCG2, CDK18, CDK2, CDK6, CDK8, CENPA, DHFRL1*. Epirubicin - *ABCB1, CDA, CYP1B1, ERBB3, ERCC1, GSTP1, MTHFR, NOS3, ODC1, PON1, RAD50, SEMA4D, TFDP2*. Doxorubicin - *ABCB1, ABCC2, ABCD3, AKR1B1, AKR1C1, CBR1, CYBA, FTH1, FTL, GPX1, MT2A, NCF4, RAC2, SLC22A16, TXNRD1*. 5-Fluorouracil - *ABCB1, ABCC3, CFLAR, IL6, MTHFR, TP53, UCK2*.

¹ Surviving patients; ² Analysis included patients in the METABRIC 'discovery' dataset only; ³ SVMs tested with 9 fold cross-validation, all others tested with leave-one-out cross-validation; ⁴ Includes all patients treated with HT/CT, combination CT/HT, either with or without combination radiotherapy; ⁵ Median time after treatment until death (> 4.4 years) was used to distinguish favorable outcome, ie. sensitivity to therapy.

paclitaxel signature in patients treated with HT and/or adjuvant chemotherapy (78.6%).

The RF classifier was used to predict paclitaxel therapy outcome for patients that underwent CT and/or HT (Table 2). The best performance achieved with RF showed 82.4% overall accuracy using a 3-year survival threshold for distinguishing therapeutic resistance vs. sensitivity.

The best overall accuracy and AUC (sensitivity and specificity) for CT/HT patients using mRMR feature selection for SVM predicting outcome of paclitaxel therapy was obtained for CT patients with 4 year survival. Outcomes for HT patients with 3 year survival were

predicted with 84% accuracy; however the specificity was lower in this group. SVM combined with mRMR further improved accuracy of feature selection and prediction of response to hormone and/or chemotherapy based on survival time than either SVM or RF alone.

While not a replication study *sensu stricto*, the initial paclitaxel gene set used for feature selection was the same as in our previous study¹. Predictions for the METABRIC patient cohort, which was independent of the previous validation set⁵, of the either same (SVM) or different ML methods (RF and SVM with mRMR) exhibited comparable or better accuracies than our previous gene signature¹.

Table 2. Results of applying RF to predict outcome of paclitaxel therapy.

Type of treatment	Survival years (as threshold)	# Patients	Final gene expression signature	Accuracy (%)	AUC ¹
Chemotherapy (CT)	3	53	<i>ABCB1, ABCB11, ABCC1, ABCC10, BAD, BBC3, BCL2, BCL2L1, BMF, CYP2C8, CYP3A4, MAP2, MAP4, MAPT, NR1I2, SLCO1B3, TUBB1, TUBB4A, TUBB4B</i>	56.6	0.534
	4			60.4	0.645
	5			58.5	0.645
Hormone therapy (HT)	3	420		82.4	0.641
	4			77.4	0.555
	5			70.7	0.648
CT and/or HT	3	504		79.6	0.564
	4			73.6	0.527
	5			63.5	0.563

¹AUC: Area under receiver operating curve; both Discovery and Validation patient datasets analyzed

Table 3. Results of mRMR feature selection for an SVM for predicting outcome of paclitaxel therapy.

Type of treatment	Survival years (as threshold)	Number of patients ¹	Gene signature selected by mRMR	Accuracy (%)	AUC
Chemotherapy (CT)	3	53	<i>BCL2, TWIST1</i>	73.6	0.724
	4		<i>ABCB11, ABCC1, BAD, BBC3, BCL2L1</i>	79.2	0.793
	5		<i>ABCB11, BAD, CYP2C8, CYP3A4, MAP2, MAPT, FGF2</i>	77.4	0.759
Hormone therapy (HT)	3	420	<i>ABCB11, ABCC1, BAD, BCL2, CYP2C8, CYP3A4, MAP4, MAPT, NR1I2, TUBB1, GBP1, OPRK1</i>	84.0	0.532
	4		<i>BBC3, MAP4, FGF2, OPRK1</i>	79.3	0.519
	5		<i>ABCC10, MAPT, TUBB1</i>	72.6	0.520
CT and/or HT	3	504	<i>ABCB11, ABCC1, BBC3, BCL2, BCL2L1, CYP2C8, CYP3A4, MAP2, MAPT, TLR6, TWIST1</i>	74.6	0.565
	4		<i>ABCB11, BAD, BCL2, CYP3A4, MAP2, MAP4, NR1I2, OPRK1, TWIST1</i>	75.0	0.535
	5		<i>TWIST1, BCL2, BMF, CYP2C8, CYP3A4, BCL2L1, BBC3, TLR6, BAD, MAP4, NR1I2, GBP1, NFKB2</i>	65.9	0.525

¹Predicted treatment responses for individual METABRIC patients using the described ML techniques are provided in Dataset 1.

These techniques are powerful tools which can be used to identify genes that may be involved in drug resistance, as well as predict patient survival after treatment. Future efforts to expand these models to other drugs may assist in suggesting preferred treatments in specific patients, with the potential impact of improving efficacy and reducing duration of therapy.

Data availability

Patient data: The METABRIC datasets are accessible from the European Genome-Phenome Archive (EGA) using the accession number EGAS00000000083 (<https://www.ebi.ac.uk/ega/studies/EGAS00000000083>). Normalized patient expression data for the discovery (EGAD00010000210) and validation sets (EGAD00010000211) were retrieved with permission from EGA. Corresponding clinical data was obtained from the literature⁶. While not individually curated, HT patients were treated with tamoxifen and/or aromatase inhibitors, while CT patients were most commonly treated with cyclophosphamide-methotrexate-fluorouracil (CMF), epirubicin-CMF, or doxorubicin-cyclophosphamide.

F1000Research: Dataset 1. Predicted treatment response for each individual METABRIC patient, [10.5256/f1000research.9417.d133983](https://doi.org/10.5256/f1000research.9417.d133983)¹⁰

Author contributions

PKR, AN and LR designed the methodology and oversaw the project. SVM feature selection with MATLAB was automated by DA. EJM and KB selected the initial gene signatures, and performed processing of the METABRIC data using SVM methods. IR performed the preprocessing of the METABRIC dataset using RF; IR and HQ designed feature selection and classification modules using WEKA. PKR, IR, EJM, AN, and LR wrote the manuscript.

Competing interests

PKR cofounded Cytognomix. A patent application related to biologically inspired gene signatures is pending. The other authors declare that they have no competing interests.

Grant information

AN and LR are funded by NSERC grants RGPIN-2016-05017 and RGPIN-2014-05084 and by the Windsor Essex County Cancer Centre Foundation under a Seeds4Hope grant. PKR has been supported by NSERC [Discovery Grant 371758-2009], Canadian Foundation for Innovation, Canada Research Chairs and Cytognomix Inc.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

- Dorman SN, Baranova K, Knoll JH, *et al.*: **Genomic signatures for paclitaxel and gemcitabine resistance in breast cancer derived by machine learning.** *Mol Oncol.* 2016; **10**(1): 85–100.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Daemen A, Griffith OL, Heiser LM, *et al.*: **Modeling precision treatment of breast cancer.** *Genome Biol.* 2013; **14**(10): R110.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Shoemaker RH: **The NCI60 human tumour cell line anticancer drug screen.** *Nat Rev Cancer.* 2006; **6**(10): 813–823.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Park NI, Rogan PK, Tarnowski HE, *et al.*: **Structural and genic characterization of stable genomic regions in breast cancer: Relevance to chemotherapy.** *Mol Oncol.* 2012; **6**(3): 347–59.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Hatzis C, Pusztai L, Valero V, *et al.*: **A genomic predictor of response and survival following taxane-anthracycline chemotherapy for invasive breast cancer.** *JAMA.* 2011; **305**(18): 1873–1881.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Curtis C, Shah SP, Chin SF, *et al.*: **The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups.** *Nature.* 2012; **486**(7403): 346–352.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- MATLAB and Statistics Toolbox Release 2014a.** The MathWorks Inc., Natick, Massachusetts, United States.
- Hall M, Frank E, Holmes G, *et al.*: **The WEKA data mining software: an update.** *ACM SIGKDD explorations newsletter.* 2009; **11**(1): 10–18.
[Publisher Full Text](#)
- Ding C, Peng H: **Minimum redundancy feature selection from microarray gene expression data.** *J Bioinform Comput Biol.* 2005; **3**(2): 185–205.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Rezaeian I, Mucaki EJ, Baranova K, *et al.*: **Dataset 1 in: Predicting Outcomes of Hormone and Chemotherapy in the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) Study by Machine Learning.** *F1000Research.* 2016.
[Data Source](#)

Open Peer Review

Current Peer Review Status: ? ?

Version 1

Reviewer Report 03 October 2016

<https://doi.org/10.5256/f1000research.10141.r16345>

© 2016 Tung C. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Chun-Wei Tung

School of Pharmacy, Kaohsiung Medical University, Kaohsiung, Taiwan

This study proposed prediction methods using SVM and RF classifiers with mRMR selected feature sets from cell line data and demonstrate its prediction ability for outcomes from METABRIC patient cohort. The classifiers with good prediction performance show the usefulness of combining domain knowledge with feature selection techniques. However, some details essential for reproducibility and interpretation are missing.

Required information is listed in the following.

1. What are the values of parameters for SVM and RF classifiers and the methods for parameter selection (by default or other selection methods)?
2. The development and evaluation of models for patient data are not clear. Whether the models were trained using partial data from METABRIC or only leave-one-out cross-validation was applied? If cross-validation is the case, then what is the model offered at the online server because there will be more than one models created, and whether the cross-validation is involved in the feature selection process that often leads to an overestimation of the performance. For the case of training on partial data, both training and test performance are essential information for evaluating the robustness of models.
3. Since some of the datasets are highly imbalanced, the numbers of positives and negatives, as well as sensitivity and specificity are more important than accuracy for interpreting the results as a high accuracy with a low AUC could be the result of all positive/negative predictions on an imbalanced dataset. Listing all the information along with the accuracy and AUC will help the interpretation of prediction performances.

Competing Interests: No competing interests were disclosed.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 13 Jan 2017

Peter Rogan

Comment 1: What are the values of parameters for SVM and RF classifiers and the methods for parameter selection (by default or other selection methods)?

Response: The parameter values for these classifiers have been added to the Tables 1-5.

In regards to parameter selection, the first paragraph of the methods now describes C and Sigma selection as a grid search to find the values with the lowest cross-validation misclassification rate. Similarly for RF, a grid search was used to optimize the maximum number of randomly selected genes for each tree (second paragraph of Methods section).

Comment 2: The development and evaluation of models for patient data are not clear. Whether the models were trained using partial data from METABRIC or only leave-one-out cross-validation was applied? If cross-validation is the case, then what is the model offered at the online server because there will be more than one models created, and whether the cross-validation is involved in the feature selection process that often leads to an overestimation of the performance. For the case of training on partial data, both training and test performance are essential information for evaluating the robustness of models.

Response: We obtained new results for both RF and mRMR+SVM models when we use discovery set as training set and validation set as test set, the performance of the model was poor. After more investigation we found that there happened to be a large variation between gene expression of 26 targeted genes between discovery and validation set (please see Supplementary Dataset 2). Hence, building any classifier using discovery and validation set as training and test set in their current forms will result of poor performance, since the training and test sets are vastly different.

However, we did carry out another experiment on discovery set solely and used 70% of data for training and remaining 30% for test the performance of the model. The results have been added to the manuscript (Tables 4 and 5).

Comment 3: Since some of the datasets are highly imbalanced, the numbers of positives and negatives, as well as sensitivity and specificity are more important than accuracy for interpreting the results as a high accuracy with a low AUC could be the result of all positive/negative predictions on an imbalanced dataset. Listing all the information along with the accuracy and AUC will help the interpretation of prediction performances.

Response: As previously mentioned, we have added more performance measures including MCC and AUC. They have been added Tables 1-5 of the manuscript.

Competing Interests: PKR cofounded Cytognomix. A patent application related to biologically inspired gene signatures is pending. The other authors declare that they have no competing interests.

Reviewer Report 30 September 2016

<https://doi.org/10.5256/f1000research.10141.r16733>

© 2016 Fertig E. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Elana J. Fertig

Division of Oncology Biostatistics and Bioinformatics, School of Medicine, Johns Hopkins University, Baltimore, MD, USA

This study develops SVM and RF algorithms built upon previously learned gene signatures of therapeutic response to breast cancer. The algorithms are applied and compared to predict patient survival under different treatment conditions in METABRIC data. The analyses and comparisons are robust and this study provides a useful assessment of biologically-driven classifiers. The three major areas that require improvement before the article is indexed are as follows, and described in further detail below.

1. The methods require further clarification to distinguish differences between this study and the previous study as well as the parameters of the machine learning algorithms.
2. Accuracy in the results must better distinguish results on independent test and training sets.
3. Classifiers must be put in the context of other existing genomics classifiers used in breast cancer and/or previously published in Mammprint data.

Title and Abstract

Acceptable

Article content

Methods

1. Abbreviations SVM and RF must be spelled out as Support Vector Machine and Random Forrest on first use in Methods.
2. Writing in *SVM learning* subsection of *Methods* requires clarification to distinguish which of these methods were developed in the previous *Molecular Oncology* publication and which were developed as part of this publications.
3. Details about the SVM learning algorithm are included in the caption to Figure 1, but must also be included and completely described in text for the corresponding section of the methods.
4. No equations are provided to describe the role of the parameters C and sigma. It is also unclear whether this greedy search is implemented by the Matlab function *fitcsvm* or uses custom code developed by the authors.

Results

1. Need to specify whether reported accuracies are computed with leave-one-out cross validation or 9-fold cross validation (described in Methods).

2. Ideally, given the size of METABRIC data they would be calculated on independent training (first 1000 patient samples) and training (last 1000 patient samples) datasets.
3. AUC must be computed separately for discovery and validation sets (Table 2).
4. It is unclear whether the previous validation set described in the sentence “Predictions for the METABRIC patient cohort, which was independent of the previous validation set” refers to a validation set used in this publication or the previous publication.
5. Covariates such as ER/PR or PAM50 subtypes must be included in a table describing the sample cohorts. Accuracy must be computed separately for these co-variables or they must also be included as co-variables in the machine-learning model.
6. Ideally accuracy would be compared to existing breast cancer classifiers (e.g., using code from Marchionni *et al.*, *BMC Genomics*, 2013) and/or survival curves reported in the literature.

Conclusions

1. Must be discussed in the context of existing genomics classifiers for breast cancer (e.g., OncotypeDx and/or Mammaprint).
2. Results must be put in context with other predictions on METABRIC data, e.g., outcomes from the DREAM contest.

Data

Acceptable

Competing Interests: No competing interests were disclosed.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 13 Jan 2017

Peter Rogan

Comment 1: The methods require further clarification to distinguish differences between this study and the previous study as well as the parameters of the machine learning algorithms.

Response: The first paragraph of the Methods describes Support Vector Machine learning, which has been greatly expanded upon. Differences in SVM methodology between the two studies are indicated there (i.e. a Gaussian kernel was used instead of a linear kernel). All other feature selection methods described in the manuscript (Random Forest, mRMR) were not used in Dorman *et al.*, 2016.

The parameters for machine learning algorithms have been incorporated in the manuscript, and can be found in the footnote section of each data table.

Comment 2: Accuracy in the results must better distinguish results on independent test and training sets.

Response: The Validation dataset showed a distinct overall expression profile from the Discovery set, possibly due to batch effects, which are well known. We added another experiment to the manuscript by splitting the Discovery set into Training and Test sets. The model was trained using 70% of the data and then tested using the remaining 30% of data as test set. We repeated this procedure 100 times and took the median as the final performance result. The results are presented in Tables 4 and 5 of the manuscript.

Comment 3: Classifiers must be put in the context of other existing genomics classifiers used in breast cancer and/or previously published in MammaPrint data.

Response: We have added two sentences in the second paragraph of the “Results and Discussion” section which describes the comparison of our gene signature to those from MammaPrint and Oncotype Dx. Pair-wise comparison of these three signatures show that they are nearly independent of one another.

Methods

Comment 4: Abbreviations SVM and RF must be spelled out as Support Vector Machine and Random Forest on first use in Methods.

Response: We thank the reviewer for this suggestion. It has been addressed in the Methods section of the manuscript.

Comment 5: Writing in SVM learning subsection of Methods requires clarification to distinguish which of these methods were developed in the previous Molecular Oncology publication and which were developed as part of this publications.

Response: This is now clarified within the first paragraph of the Methods section in the manuscript. The SVM classifier was adopted from previous Molecular Oncology publication, while the feature selection method has been developed as part of this publication.

Comment 6: Details about the SVM learning algorithm are included in the caption to Figure 1, but must also be included and completely described in text for the corresponding section of the methods.

Response: Thanks for the reviewer’s suggestion. This description of the SVM learning algorithm has been moved from the Figure 1 legend and integrated into the first paragraph of the methods section.

Comment 7: No equations are provided to describe the role of the parameters C and sigma. It is also unclear whether this greedy search is implemented by the Matlab function fitcsvm or uses custom code developed by the authors.

Response: A brief description of the role of each parameter has been added to the first paragraph of the methods section of the manuscript. Readers are also now directed to a reference (Ben-Hur and Weston, 2010) if more detail is desired.

The greedy search, also called sequential backward feature selection, was implemented as a script by our lab in MATLAB. It is not a MATLAB function. This is clarified by changing a few words in the first paragraph of the methods section: "A backwards feature selection (greedy) algorithm was designed and implemented in MATLAB in which..."

Moreover, as described above, the SVM classifier was adopted from previous Molecular Oncology publication (Dorman *et al.* 2016), while the feature selection method has been developed as part of this publication.

Results

Comment 8: Need to specify whether reported accuracies are computed with leave-one-out cross validation or 9-fold cross validation (described in Methods).

Response: All SVM models described in the manuscript used leave-one-out cross validation except one, and this is clearly indicated in Table 1, and is now commented on in the methods. A 9-fold cross-validation was used to build a model using 735 patients who were treated with Chemotherapy and/or Hormone therapy, as leave-one-out cross validation of this many patients took an unreasonably long time to complete (it exceeded 3 weeks on a dedicated I7 Intel processor).

Comment 9: Ideally, given the size of METABRIC data they would be calculated on independent training (first 1000 patient samples) and test (last 1000 patient samples) datasets.

Response: We obtained new results for both RF and mRMR+SVM models using Discovery patient set for training and Validation set for testing, however the performance of the model was poor. After further investigation, we found that there were large differences between gene expression levels of the 26 model signature genes in the Discovery versus Validation sets (we used Wilcoxon rank sum test, Kruskal-Wallis test and t-test to evaluate the results – shown in the plotted distributions of gene expression in Supplemental Dataset 2) regardless of patient status (alive or dead). Hence, building any classifier using discovery and validation set as training and test set in their current forms will result of poor performance due to this source of heterogeneity.

To address this issue, we did carry out another experiment based on data from the Discovery patient dataset alone; using 70% of data for training and remaining 30% for testing, the performance of the model was significantly better. We speculate that the discrepancy between the expression distributions in the Discovery and Validation sets were the result of batch effects. The results have been added to the manuscript (Tables 4,5).

Comment 10: AUC must be computed separately for discovery and validation sets (Table 2).

Response: We have included additional performance measures to Tables 1-5, including Area Under Curve (AUC).

Comment 11: It is unclear whether the previous validation set described in the sentence

"Predictions for the METABRIC patient cohort, which was independent of the previous validation set" refers to a validation set used in this publication or the previous publication.

Response: This sentence is referring to breast cancer patient data from Hatzis *et al.* (2013), which was used as a validation set in Dorman *et al.* (2016), not this publication. We have modified this sentence to clarify the issue.

Comment 12: Covariates such as ER/PR or PAM50 subtypes must be included in a table describing the sample cohorts. Accuracy must be computed separately for these co-variables or they must also be included as co-variables in the machine-learning model.

Response: Even with the subtype as covariant, it is not possible to perform the analysis the reviewer requested. Certain therapies are definitely more effective in particular subtypes (eg. etoposide, docetaxel, and cisplatin are preferentially active in basal or claudin-low cell lines, as observed clinically; Heiser *et al.*, 2012). The public METABRIC dataset (or the corresponding publication) does not provide the specific therapies used to treat individual patients. Had they done so, it would have made sense to look at these covariates.

Reference: Heiser LM, Sadanandam A, Kuo WL, Benz SC, Goldstein TC, Ng S, Gibb WJ, Wang NJ, Ziyad S, Tong F, *et al.* (2012). Subtype and pathway specific responses to anticancer compounds in breast cancer. *Proc Natl Acad Sci US A* **109**:2724-2729.

Comment 13: Ideally accuracy would be compared to existing breast cancer classifiers (e.g., using code from Marchionni et al., BMC Genomics, 2013) and/or survival curves reported in the literature.

Response: The proposed method has been compared against the K-TSP (Marchionni *et al.*, BMC Genomics, 2013) as per reviewer's suggestion and the results are presented in Table 6 of the manuscript.

Conclusions

Comment 14: Must be discussed in the context of existing genomics classifiers for breast cancer (e.g., OncotypeDx and/or Mammaprint).

Response: We have added text to both the second paragraph of the "Results and Discussion" paragraph and to the conclusion of the paper.

Comment 15: Results must be put in context with other predictions on METABRIC data, e.g., outcomes from the DREAM contest.

Response: An important distinction to note in regards to our methodology is that the predictions are based on the genes known to be associated with the response to specific drugs used to treat breast cancer. In the DREAM contest, the method with the highest METABRIC score (as described in Cheng *et al.*, 2013) was phenotype-based, finding signatures for molecular processes that are dysregulated in METABRIC, rather than responses to the cancer therapies themselves. While this is an interesting prediction

method, the results cannot be compared to our approach. The gene signatures that we have derived contain components of many different pathways.

Reference: Cheng WY, Ou Yang TH, Anastassiou D. Biomolecular events in cancer revealed by attractor metagenes. PLoS Comput Biol. 2013;9(2):e1002920.

Competing Interests: PKR cofounded Cytognomix. A patent application related to biologically inspired gene signatures is pending. The other authors declare that they have no competing interests.

The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com

F1000Research