Tsinghua Science and Technology

Volume 20 | Issue 5 Article 10

2015

High Accuracy Gene Signature for Chemosensitivity Prediction in Breast Cancer

Wei Hu
Department of Computer Science, Houghton College, Houghton, NY 14744, USA.

TSINGHUA SCIENCE AND TECHNOLOGY
ISSN 1007-0214 10/10 pp530-536
Volume 20, Number 5, October 2015

High Accuracy Gene Signature for Chemosensitivity Prediction in Breast Cancer

Wei Hu*

Abstract: Neoadjuvant chemotherapy for breast cancer patients with large tumor size is a necessary treatment. After this treatment patients who achieve a pathologic Complete Response (pCR) usually have a favorable prognosis than those without. Therefore, pCR is now considered as the best prognosticator for patients with neoadjuvant chemotherapy. However, not all patients can benefit from this treatment. As a result, we need to find a way to predict what kind of patients can induce pCR. Various gene signatures of chemosensitivity in breast cancer have been identified, from which such predictors can be built. Nevertheless, many of them have their prediction accuracy around 80%. As such, identifying gene signatures that could be employed to build high accuracy predictors is a prerequisite for their clinical tests and applications. Furthermore, to elucidate the importance of each individual gene in a signature is another pressing need before such signature could be tested in clinical settings. In this study, Genetic Algorithm (GA) and Sparse Logistic Regression (SLR) along with *t*-test were employed to identify one signature. It had 28 probe sets selected by GA from the top 65 probe sets that were highly overexpressed between pCR and Residual Disease (RD) and was used to build an SLR predictor of pCR (SLR-28). This predictor tested on a training set ($n=81$) and validation set ($n=52$) had very precise predictions measured by accuracy, specificity, sensitivity, positive predictive value, and negative predictive value with their corresponding *P* value all zero. Furthermore, this predictor discovered 12 important genes in the 28 probe set signature. Our findings also demonstrated that the most discriminative genes measured by SLR as a group selected by GA were not necessarily those with the smallest *P* values by *t*-test as individual genes, highlighting the ability of GA to capture the interacting genes in pCR prediction as multivariate techniques. Our gene signature produced superior performance over a signature found in one previous study with prediction accuracy 92% vs 76%, demonstrating the potential of GA and SLR in identifying robust gene signatures in chemo response prediction in breast cancer.

Key words: genetic algorithm; gene signature; breast cancer; sparse logistic regression; predictor; chemosensitivity

1 Introduction

Breast cancer is a heterogeneous disease of different molecular subtypes with distinct genetic alterations and patients could have remarkably different clinical outcome and response to various therapies. Chemotherapy uses drugs to destroy cancer cells, stop their growth, or ameliorate symptoms. Adjuvant chemotherapy for breast cancer is a treatment given after primary therapy to increase the chance of long-term survival while neoadjuvant chemotherapy is given prior to primary therapy with an objective to reduce its current size, so it can be surgically removed. Neoadjuvant chemotherapy not only facilitates the procedure of surgical extraction of a tumor but can also improve postoperative recovery for

• Wei Hu is with Department of Computer Science, Houghton College, Houghton, NY 14744, USA. E-mail: wei.hu@houghton.edu.
∗ To whom correspondence should be addressed.
Manuscript received: 2015-07-06; accepted: 2015-08-06

the patient. Because of its effectiveness, neoadjuvant chemotherapy is being evaluated in other settings such as esophageal, gastric, pancreatic, prostate, ovarian, and cervical cancers. Furthermore, neoadjuvant chemotherapy can be employed to directly assess tumor response to therapy.

In current practice, chemotherapy is applied empirically, and not all patients can benefit from it, illustrating the urgent needs for a more personalized approach in cancer treatment. It is desirable, due to its clinical significance, to have the ability to predict whether an individual patient will benefit from a specific therapy. Complete eradication of all invasive cancer from the breast and regional lymph nodes, pathologic Complete Response (pCR), is associated with both breast cancer subtype and long-term survival. Single clinical or molecular parameter, such as tumor size, histology, hormone receptor or Human Epidermal growth factor Receptor 2 (HER2) expression, and tumor grade, does not always give reliable predictions of response. With microarray data, researchers are able to identify gene expression patterns that are predictive of chemotherapy response.

In a previous study[1], t-test for unequal-variance was employed to find a signature of 31 probe sets (27 genes) from patient gene expression data with highest differentially expressed values between pCR and Residual Disease (RD). Based on this signature, a 30-probe set Diagnal Linear Discriminant Analysis (DLDA-30) classifier was constructed and selected, after comparing many other predictors, to predict pathological response to preoperative paclitaxel/FAC chemotherapy. The advantage of this predictor is its ability to identify those patients most likely to benefit from a particular treatment, the neoadjuvant chemotherapy, in this case. This predictor can recognize not all responsive patients but exclusively those that will benefit the most, as defined by attaining a pCR. Additionally, they found that clinical variables such as age, nuclear grade, and ER status were significantly associated with pCR in the training set they used to build the predictor. Other clinical studies also identified gene signatures that predict response to neoadjuvant therapy of breast cancer[2–11].

As a single variable technique, t-test analyzes one gene at a time, and as a result using t-test alone might miss the interactions between the genes. We believed that a gene signature identified by t-test could be optimized with the help of a multivariable technique such as genetic algorithm. This algorithm has the capacity to explore multiple solutions concurrently, from which interacting and informative genes could be discovered. Our study aimed to apply genetic algorithm to search for novel signatures from patient gene expression profiling and use them to develop predictors of pCR that can achieve much better predictions than the DLDA-30 found in Ref. [1].

2 Datasets and Methods

2.1 Patient cohorts and clinical information

One breast cancer patient cohort was obtained from a previous publication[1] ($n = 133$). Needle-biopsy samples were collected from 133 patients with stage I, II, or III breast cancer who received preoperative weekly paclitaxel and a combination of fluorouracil, doxorubicin, and cyclophosphamide (T/FAC). These 133 patients were divided into two subsets, one training set of size 81 and one validation set of size 52. These data contain clinical information including patient age, gender, race, histological classification, stage, nuclear grade, ER (estrogen receptor), PR (progesterone receptor), and HER2 (human epidermal growth factor 2) status, pathologic complete response, and residual disease. These data also contain each patient's genome-scale gene expression profiles generated using Affymetrix U133A chip (Santa Clara, CA). pCR was defined as no residual invasive cancer in the breast or lymph nodes. pCR is presently accepted as a reasonable early indicator for long-term survival.

2.2 Top 65 probe sets

Since there are so many genes involved in one gene expression experiment, we had to narrow our search space by using the t-test to find a pool of good candidate genes for our signature. For this purpose, t-tests for unequal variances for all the probe sets on the Affymetrix U133A chip were carried out to find the genes that were significantly differentially expressed in either the pCR cases or the RD cases. We chose the 60 probe sets with the smallest t-test P values (False Discovery Rate (FDR) = 1%) and 5 probe sets with the most negative t-test statistics in the remaining probe sets to be our first signature (top 65-probe set signature) as presented in Table 1. The top 31 probe set signature in Ref. [1] had an FDR = 0.5%.

Table 1 Outline of a genetic algorithm.

Generate an initial population of individuals
Evaluate initial population
Repeat
Perform selection
Apply genetic operations such as mutation and crossover to generate a new generation of individuals
Evaluate individuals in the population
Until some stopping criterion is satisfied

2.3 Sparse logistic regression

A standard least squares linear regression solves the following problem:

Given data $\{x_i, y_i\}_{i=1}^m$, find $\alpha = (\alpha_0, \alpha_1, \alpha_2, \cdots, \alpha_n)$ such that

$$\min_{\alpha} \sum_i (y_i - f(x_i))^2 \text{ where } f(x_i) = \sum_{j=1}^{n} \alpha_j x_{ij} + \alpha_0.$$

The Least Absolute Shrinkage and Selection Operator (LASSO) regression[12], a shrinkage and selection method for linear regression, deals with the following problem:

$$\min_{\alpha} \sum_i (y_i - f(x_i))^2 \text{ with} \sum_{j=1}^{n} |\alpha_j| \leqslant t,$$

where t controls the L_1 norm of $(\alpha_1, \alpha_2, \cdots, \alpha_n)$. This constraint on α produces a sparse model, i.e., many components of α can be zero, and controls the complexity of the model. Based on the idea of LASSO, the work in Ref. [13] studied the following problem for sparse logistic regression, when $y_i \in \{-1, +1\}$,

$$\min_{\alpha} \sum_i g(-y_i f(x_i)) \text{ with } \sum_{j=1}^{n} |\alpha_j| \leqslant t,$$

where $g(\xi) = \log(1 + e^{\xi})$, which is the negative log-likelihood function associated with the probability model

$$\text{Prob}(y|x) = \frac{1}{1 + e^{-y \cdot f(x)}},$$

where the best value of the parameter t must be chosen by the user or alternatively optimized in an additional model selection stage. Subsequently, in Ref. [14], a novel technique was proposed to solve this sparse logistic regression problem efficiently with Bayesian regularization using a Laplace prior. In their method, the value of this parameter was found via a minimization of the leave-one-out cross-validation estimate of the cross-entropy loss. To employ the sparse logistic regression in our study of the gene expression data, we used +1 to label those cases of RD status, and used −1 to label those cases of pCR status.

2.4 Genetic Algorithm (GA)

Genetic algorithm, a particular kind of evolutionary algorithms, is a search algorithm that adopts some common processes in genetics such as selection, mutation, and inheritance. This algorithm outperforms other traditional search algorithms in various applications.

2.5 Statistical significance of pCR prediction

In order to evaluate the statistical significance of our predictions, we need to compare them with random predictions. For each dataset, a random-label permutation was conducted while keeping the number of instances in each group fixed. The matches between the permuted labels and the original ones were recorded. The standard P value was the percentage of 1000 random predictions with higher accuracy than the calculated predictions.

3 Results

To select informative genes for response prediction, we first identified a set of genes with t-test as shown in Table 2, some of which are shared with those found in Ref. [1]. Here we applied genetic algorithm to discover a gene signature that has high prediction accuracy as an improvement of the earlier work[1]. Finally, sparse logistic regression was employed to this signature to find the importance of each individual gene's contribution to the prediction of pCR.

3.1 Gene signature: The 28 probe sets

To use GA to search and select a subset of genes that had a high prediction accuracy from the 65 probe sets in Table 2, we represented our solution, referred to as an individual in GA terms, as a binary vector of size 65 to indicate the presence (1) or absence (0) of each probe set in the 65 probe sets. We ran the GA with population size 200, individual size 65, and 100 generations. In each generation, the top 50% of the individuals with highest fitness values were selected as

Table 2 Top 65 differentially expressed probe sets by unequal-variance *t*-test ($n = 82$, probe sets with a * are contained in the top 31 probe sets found in Ref. [1]).

Rank by *P* value	*t*-test	*P* value	Higher expression in	Probe set ID	Gene symbol	Gene name
1	6.215 265	2.20×10^{-8}	RD	203930_s_at*	MAPT	Microtubule-associated protein tau
2	6.367 41	2.31×10^{-8}	RD	203929_s_at*	MAPT	Microtubule-associated protein tau
3	6.212 778	2.56×10^{-8}	RD	212207_at*	THRAP2	Thyroid hormone receptor associated protein 2
4	5.804 489	1.25×10^{-7}	RD	212745_s_at*	BBS4	Bardet-Biedl syndrome 4
5	5.847 627	1.42×10^{-7}	RD	203928_x_at*	MAPT	Microtubule-associated protein tau
6	5.763 819	1.67×10^{-7}	RD	208945_s_at*	BECN1	Beclin 1 (coiled-coil, myosin-like BCL2 interacting protein)
7	5.704 523	2.50×10^{-7}	RD	206401_s_at*	MAPT	Microtubule-associated protein tau
8	5.716 982	2.77×10^{-7}	RD	205354_at*	GAMT	Guanidinoacetate N-methyltransferase
9	5.555 817	3.65×10^{-7}	RD	219741_x_at*	ZNF552	Zinc finger protein 552
10	5.523 853	4.08×10^{-7}	RD	215304_at*	—	Clone 23948 mRNA sequence
11	5.449 088	5.45×10^{-7}	RD	209173_at	AGR2	Anterior gradient 2 homolog (Xenopus laevis)
12	5.391 683	6.89×10^{-7}	RD	201508_at*	IGFBP4	Insulin-like growth factor binding protein 4
13	5.357 545	8.43×10^{-7}	RD	217542_at*	MDM2	Mdm2, transformed 3T3 cell double minute 2, p53 binding protein (mouse)
14	5.312 123	1.30×10^{-6}	RD	219044_at*	FLJ10916	Hypothetical protein FLJ10916
15	5.267 37	1.41×10^{-6}	RD	215616_s_at*	JMJD2B	Jumonji domain containing 2B
16	5.215 414	1.41×10^{-6}	RD	204509_at*	CA12	Carbonic anhydrase XII
17	5.221 534	1.42×10^{-6}	RD	202204_s_at*	AMFR	Autocrine motility factor receptor
18	5.215 809	1.70×10^{-6}	RD	214124_x_at*	FGFR1OP	FGFR1 oncogene partner
19	5.210 207	1.71×10^{-6}	RD	219051_x_at*	METRN	Meteorin, glial cell differentiation regulator
20	5.194 077	1.97×10^{-6}	RD	209696_at	FBP1	Fructose-1,6-bisphosphatase 1
21	5.052 227	2.70×10^{-6}	RD	213234_at*	KIAA1467	KIAA1467 protein
22	5.049 412	2.74×10^{-6}	RD	217838_s_at	EVL	Enah/Vasp-like
23	5.054 632	2.95×10^{-6}	RD	205074_at	SLC22A5	Solute carrier family 22 (organic cation transporter), member 5
24	5.139 071	3.06×10^{-6}	RD	213623_at*	KIF3A	Kinesin family member 3A
25	5.017 933	3.38×10^{-6}	RD	201413_at	HSD17B4	Hydroxysteroid (17-beta) dehydrogenase 4
26	4.908 014	5.26×10^{-6}	RD	205225_at	ESR1	Estrogen receptor 1
27	4.823 788	7.04×10^{-6}	RD	217016_x_at	FLJ23172	Hypothetical LOC389177
28	4.807 25	7.18×10^{-6}	RD	214053_at*	—	CDNA FLJ44318 fis, clone TRACH3000780
29	4.888 899	7.30×10^{-6}	RD	213527_s_at	ZNF688	Zinc finger protein 688
30	4.819 068	7.44×10^{-6}	RD	203009_at	LU	Lutheran blood group (Auberger b antigen included)
31	4.865 888	9.07×10^{-6}	RD	212046_x_at	MAPK3	Mitogen-activated protein kinase 3
32	4.854 113	9.27×10^{-6}	RD	205012_s_at	HAGH	Hydroxyacylglutathione hydrolase
33	4.762 182	9.56×10^{-6}	RD	203675_at	NUCB2	Nucleobindin 2
34	4.700 102	1.07×10^{-5}	RD	203071_at	SEMA3B	Sema domain, immunoglobulin domain (Ig), short basic domain, secreted, (semaphorin) 3B
35	4.710 655	1.07×10^{-5}	RD	210129_s_at	TTLL3	Tubulin tyrosine ligase-like family, member 3
36	4.671 287	1.20×10^{-5}	RD	218671_s_at	ATPIF1	ATPase inhibitory factor 1
37	4.689 638	1.23×10^{-5}	RD	209339_at	SIAH2	Seven in absentia homolog 2 (Drosophila)
38	4.629 403	1.44×10^{-5}	RD	218976_at	DNAJC12	DnaJ (Hsp40) homolog, subfamily C, member 12
39	4.649 829	1.44×10^{-5}	RD	205734_s_at	AFF3	AF4/FMR2 family, member 3
40	4.634 054	1.65×10^{-5}	RD	202641_at	ARL3	ADP-ribosylation factor-like 3
41	4.580 441	1.68×10^{-5}	RD	218259_at	MKL2	MKL/myocardin-like 2
42	4.590 716	1.71×10^{-5}	RD	220540_at	KCNK15	Potassium channel, subfamily K, member 15
43	4.578 743	1.71×10^{-5}	RD	210831_s_at	PTGER3	Prostaglandin E receptor 3 (subtype EP3)
44	4.608 731	1.77×10^{-5}	RD	218769_s_at	ANKRA2	Ankyrin repeat, family A (RFXANK-like), 2
45	4.587 999	1.81×10^{-5}	RD	218394_at	FLJ22386	Leucine zipper domain protein
46	4.568 723	1.82×10^{-5}	RD	216835_s_at	DOK1	Docking protein 1, 62kDa (downstream of tyrosine kinase 1)
47	4.606 517	1.98×10^{-5}	RD	221728_x_at	XIST	X (inactive)-specific transcript
48	4.582 593	2.04×10^{-5}	RD	212956_at	KIAA0882	KIAA0882 protein
49	4.531 619	2.06×10^{-5}	RD	212239_at	PIK3R1	Phosphoinositide-3-kinase, regulatory subunit 1 (p85 alpha)
50	4.521 411	2.13×10^{-5}	RD	212209_at	THRAP2	Thyroid hormone receptor associated protein 2
51	4.509 765	2.22×10^{-5}	RD	204792_s_at	WDTC2	WD and tetratricopeptide repeats 2
52	4.593 663	2.45×10^{-5}	RD	204862_s_at	NME3	Non-metastatic cells 3, protein expressed in
53	4.478 137	2.49×10^{-5}	RD	206418_at	NOX1	NADPH oxidase 1
54	4.538 231	2.74×10^{-5}	RD	205059_s_at	IDUA	Iduronidase, alpha-L-
55	4.463 108	2.74×10^{-5}	RD	210958_s_at	MAST4	Microtubule associated serine/threonine kinase family member 4
56	4.501 318	2.76×10^{-5}	RD	202228_s_at	SDFR1	stromal cell derived factor receptor 1
57	4.539 226	2.83×10^{-5}	RD	212660_at	PHF15	PHD finger protein 15
58	–5.016 05	2.96×10^{-5}	pCR	213134_x_at*	BTG3	BTG family, member 3
59	4.427 751	2.98×10^{-5}	RD	203789_s_at	SEMA3C	Sema domain, immunoglobulin domain (Ig), short basic domain, secreted, (semaphorin) 3C
60	4.484 732	3.00×10^{-5}	RD	216109_at	THRAP2	Thyroid hormone receptor associated protein 2
61	–5.015 38	3.31×10^{-5}	pCR	205548_s_at*	BTG3	BTG family, member 3
62	–4.531 99	0.000 122	PCR	204825_at*	MELK	Maternal embryonic leucine zipper kinase
63	–4.003 15	0.000 496	pCR	205339_at	SIL	TAL1 (SCL) interrupting locus
64	–3.977 77	0.000 442	PCR	203693_s_at*	E2F3	E2F transcription factor 3
65	–3.946 34	0.000 361	PCR	216237_s_at	MCM5	MCM5 minichromosome maintenance deficient 5, cell division cycle 46 (S. cerevisiae)

parents to produce the next generation with crossover and a point mutation was applied to each individual randomly at six genes. In each generation of GA, we divided the training set ($n = 82$) into five equal subsets and used four subsets as a training set for SLR and one subset as a test set to get the accuracy of SLR on this test set. The fitness value is the average of the prediction accuracy on the five test sets. Our goal was to choose an individual that has a higher accuracy than DLDA-30 on the training set through this five-fold cross validation process. We found an individual of this quality and its binary representation has 28 ones, which was our signature (Table 2).

3.2 The pCR predictor: SLR-28

In Ref. [1], the DLDA-30 was selected as the best predictor after a thorough search of different predictors based on Support Vector Machine (SVM), Diagonal Linear Discrimmant Analysis (DLDA), and K-Nearest Neighbor (KNN). We developed one SLR-based predictor of high precision, which used the 28 probe sets (SLR-28). The evaluation of prediction performance was conducted on both the training set ($n = 82$) and the validation set ($n = 51$). Since the DLDA-30 was evaluated on the training set with five-fold cross validation, we performed five-fold evaluation as well for our predictor. Further, we repeated the five-fold cross validation 10 times and the averaged results were reported in Table 3.

On the training set, our predictors SLR-28 produced much better predictions across all five measurements than DLDA-30 (Table 3). In Ref. [1], authors calculated the P values of their DLDA-30 predictions on the training set in five-fold cross validation and they were all zero. Since our two predictors had higher values in all the five measurements, we concluded that they also have P value zero in all these five measurements.

On the validation set, SLR-28 trained on the training set had a much higher accuracy, a much higher specificity, and a much higher PPV than DLDA-30 (Table 4). SLR-28 had P values zero in all five measurements, whereas DLDA-30 had three P values larger than 0.05, especially those for accuracy and specificity. SLR-28 correctly identified all but two who achieved pCR and all but two who achieved RD (Table 5). Tables 3 and 4 show that SLR-28 can predicts pCR with high precision on the training set and the validation set.

Table 3 Prediction measures (five-fold cross validation) of DLDA-30 and SLR-28 on the training set with all P values zero in the five measurements.

	Accuracy	Sensitivity	Specificity	PPV	NPV
DLDA-30	0.83	0.75	0.73	0.50	0.90
SLR-28	0.90	0.90	0.89	0.75	0.96

Table 4 Prediction measures of DLDA-30 and SLR-28 on the validation set along with their P values (in parentheses).

	Accuracy	Sensitivity	Specificity	PPV	NPV
DLDA-30	0.76 (0.2)	0.92 (0)	0.71 (1)	0.52 (0.1)	0.96 (0)
SLR-28	0.92 (0)	0.85 (0)	0.95 (0)	0.85 (0)	0.95 (0)

3.3 Important genes in the signature

SLR-28 also identified the important genes in the gene signature that had nonzero SLR weights (Fig. 1). The genes with zero weight did not contribute to the prediction. The genes with positive weight contribute positively to the RD prediction and those with negative weight contribute positively to pCR prediction. In most cases, genes with positive t-test statistic had positive weight like the three genes in Fig. 1, BGT3, MELK, and MCM5. However, there were some exceptions. Two genes, STUB1 and PDPK1, had positive t-test statistic, but negative weight in Table 2, demonstrating that GA as a multivariable technique could capture interactions between genes whereas t-test may miss. Our findings showed that the most discriminative genes measured by SLR as a group selected by GA were not necessarily those with the smallest P values by t-test as individual genes.

One of the important genes in Fig. 1 is CA12, which is a membrane zinc metalloenzyme that is present in different normal tissues but is overexpressed in some cancers such as renal cell and breast cancers. Two studies found that increased CA IX expression is

Table 5 Confusion matrices for DLDA-30 and SLR-28 on the validation set.

DLDA-30	Predicted as pCR	Predicted as RD	SLR-28	Predicted as pCR	Predicted as RD
Observed pCR	12	1	Observed pCR	11	2
Observed RD	11	27	Observed RD	2	36

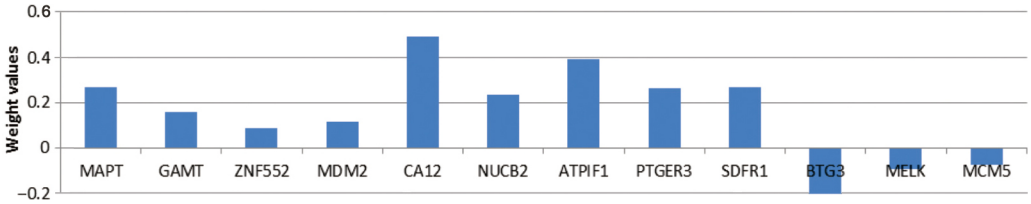


Fig. 1 Important genes in the 28 probe set signature measured by SLR.

associated with poor relapse free and overall survival in invasive breast cancer[15, 16]. Another recent study found that CA12 is regulated by estrogen receptor in breastcancer, and that this regulation involves a distal estrogen-responsive enhancer region[17].

The mitochondrial ATPase Inhibitory Factor 1 (ATPIF1) is another important gene in Fig. 1. Several studies discovered that this gene has an important role in metabolic and genetic changes observed during malignant growth[18, 19]. The large positive weight of this gene in Fig. 1 is also in agreement with this observation.

Studies in Ref. [20] revealed that MAPT is the best single gene discriminator of pCR to preoperative chemotherapy with paclitaxel, 5-fluorouracil, doxorubicin, and cyclophosphamide. There are four probe sets of MAPT selected in Table 2 and MAPT is one of the 16 informative genes in the top 65 probe sets in Fig. 1. Further, SDFR1 encodes a cell surface protein of the immunglobulin superfamily that regulates cell adhesion and process outgrowth. Finally, BTG3 is a tumor suppressor[21] and its large negative weight implies that its presence enhances chemo sensitivity.

4 Conclusions

Neoadjuvant chemotherapy is usually offered to breast cancer patients to shrink the tumor and any involved lymph nodes. Pathologic complete response is a way of measuring response to this treatment, which is normally defined as the absence of residual invasive disease in the breast and in the axillary lymph nodes at the completion of the treatment. Clinically patients who achieve pCR tend to have a very good prognosis.

As a standard treatment for breast cancer, neoadjuvant chemotherapy offers several benefits to patients. But not all patients benefit equally from this treatment. Therefore it is of great value to investigate what molecular information obtained from a patient's tumor could be used to determine if the patient would benefit from a particular chemotherapy. In practice the estrogen receptor status could be used to guide the decisions on hormonal therapy. Further, the gene expression data that reflect subtle differences in tumors could be utilized to build a predictor of response to cancer drugs.

In this study we sought to use gene expression data to predict who might achieve pCR to sequential anthracycline paclitaxel preoperative chemotherapy. Our aim was to uncover one gene signature for developing predictors of pCR to neoadjuvant chemotherapy that have a much higher accuracy than DLDA-30, so it might have the potential for clinical applications. With the ability to account for multiple gene interactions, the multivariable techniques, such as genetic algorithm and SLR, have demonstrated their utility in identifying robust gene signatures of clinical relevance.

References

[1] K. R. Hess, K. Anderson, W. F. Symmans, V. Valero, N. Ibrahim, J. A. Mejia, D. Booser, R. L. Theriault, A. U. Buzdar, P. J. Dempsey, et al., Pharmacogenomic predictor of sensitivity to preoperative chemotherapy with paclitaxel and fluorouracil, doxorubicin, and cyclophosphamide in breast cancer, *J. Clin. Oncol*., vol. 24, pp. 4236–4244, 2006.

[2] M. Chanrion, V. Negre, H. Fontaine, N. Salvetat, F. Bibeau, G. Mac Grogan, L. Mauriac, D. Katsaros, F. Molina, C. Theillet, et al., A gene expression signature that can predict the recurrence of tamoxifen-treated primary breast cancer, *Clin. Cancer Res.*, vol. 14, no. 6, pp. 1744–1752, 2008.

[3] S. P. Linke, T. M. Bremer, C. D. Herold, G. Sauter, and C. Diamond, A multimarker model to predict outcome in tamoxifen-treated breast cancer patients, *Clin. Cancer Res*., vol. 12, no. 4, pp. 1175–1183, 2006

[4] P. E. Lønning, S. Knappskog, V. Staalesen, R. Chrisanthar, and J. R. Lillehaug, Breast cancer prognostication and prediction in the postgenomic era, *Annals of Oncology*, vol. 18, pp. 1293–1306, 2007

[5] M. A. Folgueira, D. M. Carraro, H. Brentani, D. F. Patrão, E. M. Barbosa, M. M. Netto, J. R. Caldeira, M. L. Katayama, F. A. Soares, C. T. Oliveira, et al., Gene expression profile associated with response to doxorubicin-based therapy in breast cancer, *Clin. Cancer Res.*, vol. 11, no. 20, pp. 7434–7443, 2005.

[6] H. K. Dressman, C. Hans, A. Bild, J. A. Olson, E. Rosen, P. K. Marcom, V. B. Liotcheva, E. L. Jones, Z. Vujaskovic, J. Marks, et al., Gene expression profiles of multiple breast cancer phenotypes and response to neoadjuvant chemotherapy, *Clinical Cancer Research*, vol. 12, pp. 819–826, 2006

[7] O. Thuerigen, A. Schneeweiss, G. Toedt, P. Warnat, M. Hahn, H. Kramer, B. Brors, C. Rudlowski, A. Benner, F. Schuetz, et al., Gene expression signature predicting pathologic complete response with gemcitabine, epirubicin, and docetaxel in primary breast cancer, *Journal of Clinical Oncology*, vol. 24, no. 12, pp. 1839–1845, 2006.

[8] N. S. Goldstein, D. Decker, D. Severson, S. Schell, F. Vinici, J. Margolis, and N. S. Dekhne, Molecular classification system identifies invasive breast carcinoma patients who are most likely and those who are least likely to achieve a complete pathologic response after neoadjuvant chemotherapy, *Cancer*, vol. 110, pp. 1687–1696, 2007.

[9] J. C. Chang, E. C. Wooten, A. Tsimelzon, S. G. Hilsenbeck, M. C. Gutierrez, R. Elledge, S. Mohsin, C. K. Osborne, G. C. Chamness, D. Craig Allred, et al., Gene expression profiling for the prediction of therapeutic response to docetaxel in patients with breast cancer, *Lancet*, vol. 362, pp. 362–369, 2003

[10] L. Gianni, M. Zambetti, K. Clark, J. Baker, M. Cronin, J. Wu, G. Mariani, J. Rodriguez, M. Carcangiu, D. Watson, et al., Gene expression profiles in paraffin-embedded core biopsy tissue predict response to chemotherapy in women with locally advanced breast cancer, *J. Clin. Oncol*., vol. 23, pp. 7265–7277, 2005.

[11] J. Hannemann, H. M. Oosterkamp, C. A. J. Bosch, A. Velds, L. F. A. Wessels, C. Loo, E. J. Rutgers, S. Rodenhuis, and M. J. van de Vijver, Changes in gene expression associated with response to neoadjuvant chemotherapy in breast cancer, *J. Clin. Oncol*., vol. 23, pp. 3331–3342, 2005.

[12] R. Tibshirani, Regression shrinkage and selection via the lasso, *J. Royal Statist. Soc. B*, vol. 58, pp. 267–288, 1996.

[13] S. K. Shevade and S. S. Keerthi, A simple and efficient algorithm for gene selection using sparse logistic regression, *Bioinformatics*, vol. 19, pp. 2246–2253, 2003.

[14] G. C. Cawley and L. C. Talbot, Gene selection in cancer classification using sparse logistic regression with Bayesian regularization, *Bioinformatics*, vol. 22, pp. 2348–2355, 2006.

[15] S. K. Chia, C. C. Wykoff, P. H. Watson, C. Han, R. D. Leek, J. Pastorek, K. C. Gatter, P. Ratcliffe, and A. L. Harris, Prognostic significance of a novel hypoxia-regulated marker, carbonic anhydrase IX, in invasive breast carcinoma, *J. Clin. Oncol.*, vol. 19, no. 16, pp. 3660–3668, 2001.

[16] S. A. Hussain, R. Ganesan, G. Reynolds, L. Gross, A. Stevens, J. Pastorek, P. G. Murray, B. Perunovic, M. S. Anwar, L. Billingham, et al., Hypoxia-regulated carbonic anhydrase IX expression is associated with poor survival in patients with invasive breast cancer, *Br. J. Cancer*, vol. 96, no. 1, pp. 104–109, 2007.

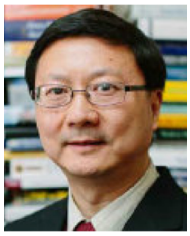
[17] D. H. Barnett, S. Sheng, T. H. Charn, A. Waheed, W. S. Sly, C.-Y. Lin, E. T. Liu, and B. S. Katzenellenbogen, Estrogen receptor regulation of carbonic anhydrase XII through a distal enhancer in breast cancer, *Cancer Research*, vol. 68, no. 9, pp. 3505–3515, 2008.

[18] M. C. Abba, Y. Hu, C. C. Levy, S. Gaddis, F. S. Kittrell, Y. Zhang, J. Hill, R. P. Bissonnette, D. Medina, P. H. Brown, and C. M. Aldaz, Transcriptomic signature of Bexarotene (Rexinoid LGD1069) on mammary gland from three transgenic mouse mammary cancer models, *BMC Medical Genomics*, vol. 1, p. 40, 2008

[19] A. Isidoro, E. Casado, A. Redondo, P. Acebo, E. Espinosa, A. M. Alonso, P. Cejas, D. Hardisson, J. A. Fresno Vara, C. Belda-Iniesta, et al., Breast carcinomas fulfill the Warburg hypothesis and provide metabolic markers of cancer prognosis, *Carcinogenesis*, vol. 26, no. 12, pp. 2095–2104, 2005

[20] R. Rouzier, R. Rajan, K. R. Hess, D. Gold, J. Stec, and M. Ayers, Microtubule associated protein tau is a predictive marker and modulator of response to paclitaxel-containing preoperative chemotherapy in breast cancer, *Proc. Natl. Acad. Sci. USA*, vol. 102, pp. 8315–8320, 2005.

[21] S. Majid, A. A. Dar, A. E. Ahmad, H. Hirata, K. Kawakami, V. Shahryari, S. Saini, Y. Tanaka, A. V. Dahiya, G. Khatri, et al., BTG3 tumor suppressor gene promoter demethylation, histone modification and cell cycle arrest by genistein in renal cancer, *Carcinogenesis*, vol. 30, no. 4, pp. 662–670, 2009.



Wei Hu is a professor of math and computer science at Houghton College. He graduated from the University of Kentucky in 1997 with a PhD degree in math and an MS degree in computer science. His research interests are data mining and machine learning.