The International Journal of Biological Markers, Vol. 22 no. 1, 2007, pp. 1-11

Individualized survival and treatment response predictions for breast cancers using phospho-EGFR, phospho-ER, phospho-HER2/neu, phospho-IGF-IR/In, phospho-MAPK, and phospho-p70S6K proteins

L. Guo[1], J. Abraham[2], D.C. Flynn[3], V. Castranova[4], X. Shi[4], Y. Qian[4]

[1] MBR Cancer Center/Department of Community Medicine, West Virginia University, Morgantown
[2] Department of Medicine and Division of Hematology/Oncology, West Virginia University, Morgantown
[3] MBR Cancer Center/Department of Microbiology, Immunology, and Cell Biology, School of Medicine, West Virginia University, Morgantown
[4] The Pathology and Physiology Research Branch, Health Effects Laboratory Division, National Institute for Occupational Safety and Health, Morgantown, West Virginia - USA

ABSTRACT: The development and progression of breast cancer involves the activation of numerous protein kinases, and the change in phosphorylation is a hallmark of protein kinase activation. In this study, we identified a comprehensive profile to predict individual breast cancer patients' survival and treatment responses using the *Random Committee* algorithm. The profile incorporated a subset of phosphorylated signal protein expressions and several selected clinical factors of breast cancer. The parameters of our profile were identified by supervised feature selection algorithms, *Gain Ratio Attribute Evaluation and Relief*. The results showed that the overall accuracy of survival prediction reached 92.3% for individual breast cancer patients with the use of the expression profiles of phospho-EGFR, phospho-ER, phospho-HER2/neu, phospho-IGF-IR/In, phospho-MAPK, and phospho-p70^{S6K} plus the selected clinical factors. The results also indicated that the overall accuracy of treatment response prediction was 92.6% with the use of the level of phospho-EGFR, phospho-ER, phospho-HER2/neu, phospho-MAPK, and phospho-p70^{S6K} plus the selected clinical information. The prediction system combines multiple signal protein activation profiles and relevant clinical information, and provides a unique guideline to aid individualized decision-making in the clinical management of breast cancer. (Int J Biol Markers 2007; 22: 1-11)

Key words: Breast cancer, Prediction, Survival, Treatment responses, Active protein kinase

INTRODUCTION

Breast cancer is a complex and heterogeneous disease encompassing a wide variety of pathological entities, clinical behaviors, and molecular profiles. Successful treatment for a given breast cancer patient relies on accurately predicting a patient's risk of developing metastases as well as the response to treatments (1). Substantial efforts have been made to create predictive factors for breast cancer. With the advance of knowledge in modern molecular biology and cell biology, many new predictive factors have been created by using genetic profiling and proteomic profiling.

Genetic profiling applies high-performance screening techniques, DNA microarray, to analyze breast cancer gene-expression profiles, which yield both prognostic and predictive information. However, the pattern of gene expression does not necessarily correlate with the pattern of protein expression. Genetic profiling can only reveal breast cancer information at the mRNA level. It is the protein that ultimately plays an essential role in breast cancer development and progression. It has been well-established that the route from mRNA to protein involves several biological processes, namely translation and posttranslational modification.

Even with the detection of total protein expression, this is still not sufficient to reflect the molecular mechanisms of breast cancer in vivo. The development and progression of breast cancer involves the activation of protein kinases (2). One of the characteristics of protein kinase activation is protein phosphorylation, and the dynamic biochemical processes of protein kinase phospho-

0393-6155/001-11$15.00/0

rylation and dephosphorylation are the essential mechanisms by which protein kinases conduct cell signaling transduction. Furthermore, numerous protein kinases are involved in the pathogenesis of breast cancers at multiple levels. Therefore, the simultaneous measurement of expression patterns of multiple phosphorylated protein kinases would more accurately reflect the pathogenesis of breast cancers and lead to identifying more powerful prognostic and predictive factors for breast cancers. In this study, several major breast cancer–related protein kinases including EGFR family members, ER, and IGF-IR/In were explored for their roles in individual survival and treatment response prediction. Furthermore, this study also included their 2 main downstream protein kinases, MAPK and $p70^{S6K}$.

Highly accurate prediction of breast cancer outcomes is essential for individualized decision-making in clinical care of patients towards appropriate personalized medicine and improved survival after therapy. Reliable prediction of breast cancer outcomes depends on appropriate computational algorithms. In this study, we propose a machine learning model system which consists of 2 feature selection algorithms, *Relief* and *Gain Ratio Attribute Evaluation,* and a classification method *Random Committee* (3). The feature selection algorithms enabled the identification of a comprehensive profile composed of important clinical information and activated protein kinase expression profiles for breast cancer outcome prediction. The classification method *Random Committee* accurately predicted the survival and treatment responses of individual breast cancer patients based on the identified comprehensive profiles. Together, this model system identified important activated protein kinase subsets and relevant clinical information for predicting breast cancer outcomes. It significantly increased ($p<0.05$) the overall accuracy of individualized breast cancer outcome predictions to above 92% and helped to reveal the interactions among the corresponding signal pathways in breast cancer.

MATERIALS AND METHODS

Data sources

Data for this analysis were obtained from the Biorepository of Clinomics Biosciences Inc, which is a large collection of highly characterized human tissue samples. These samples span a wide range of common diseases, including many forms and stages of cancer, neurological disorders and heart disease. Clinomics has pioneered the development of an emerging new technology known as *Tissue Microarrays* to enable researchers to simultaneously study hundreds of individual tissue samples in parallel, establishing the relative levels of protein expression in those samples and allowing conclusions as to the relevance of these proteins to disease to be made (4, 5).

Study cohort

Information of a breast cancer cohort was extracted from the Cell Signaling Database of Clinomics Biosciences Inc. The study cohort contained 269 breast tumor samples obtained from surgery. The patients had an average age of 63.7 years (ranging from 21 to 95 years). There were 24.9% with stage I, 45.0% with stage II, 17.8% with stage III, and 12.3% with stage IV breast cancer. 56.9% of patients underwent lumpectomy and node dissection, and 43.1% underwent mastectomy and node dissection. Two hundred and two (75.1%) patients accepted CMF chemotherapy. One hundred and fifty-three (56.9%) patients went through localized (breast) radiation therapy. Among 33 patients who developed metastases, 10 patients were responders to the treatments, while 23 were nonresponders. Treatment response was defined according to RECIS (6). Responders included patients with complete response (CR) and partial response (PR), and nonresponders included patients with stable disease (SD) and progressive disease (PR). Among the patients with stage 0 to III, 191 patients survived a 5-year disease-free interval, while 44 had recurrences within 5 years. The remaining patients' survival information was censored (details in Table I).

Immunohistochemistry

Immunohistochemistry methods were described previously (7, 8). Briefly, tissue specimens were treated with 3% H_2O_2 to quench endogenous peroxidase activity, followed by washing with PBS. After washing, the tissue specimens were first incubated with a specific primary antibody and then with a biotinylated secondary antibody. Substrate-Chromogen (SIGMA, St. Louis, MO, USA) was applied to the specimens according to the manufacturer's instructions, followed by staining with hematoxylin. The stain was semiquantitatively examined by pathologists (Clinomics BioSciences, Inc.) using the Allred 8-unit system (9). Staining was scored on a 0 to 5 scale, with 0 = no staining. Grades of 1 to 5 represent increased intensity of staining with 5 being strong, dark brown staining. For each tumor, represented by 1 slide, the tumor epithelial cells proportion score and intensity score were determined. Peritumoral inflammatory and stromal cells were not included in the evaluation. The proportion score included the fraction of positively stained tumor cells and was as follows: 0 = none, 1 = <1/100th; 2 = 1/100th to 1/10th; 3 = 1/10th to 1/3; 4 = 1/3 to 2/3; 5 = >2/3. The estimated average staining intensity of the positive tumor cells was expressed as follows: 0 = none; 1 = weak; 2 = intermediate; 3 = strong (9). Each protein was measured with 6 parameters: Cyto-

TABLE I - DESCRIPTION OF PATIENT INFORMATION IN STUDY COHORT (N=269)

Clinical information	Value	Occurrences
Histology	Infiltrating ductal carcinoma	214
	Lobular carcinoma	17
	Medullary carcinoma	11
	Papillary carcinoma	8
	Scirrhous invasive ductal carcinoma	13
	Tubular carcinoma	6
Surgery procedure	Lumpectomy + node dissection	153
	Mastectomy + node dissection	116
Age	21-50	96
	51-95	184
Stage (AJCC) [10]	I	67
	II	121
	III	48
	IV	33
Chemotherapy	CMF	202
	None	67
Radiation	Breast	153
	None	116
ER	Negative	91
	Positive	178
PR	Negative	80
	Positive	189
HER2/neu	Negative	226
	Positive	43
Metastasis site	Non-axillary lymph node	18
	Liver	7
	Lung	4
	Bone	3
	Brain	1
	None	236
Smoking	No	180
	Yes (>20 packs/year)	45
	Yes (2 packs/day)	7
	Yes (3 packs/day)	37
pT (AJCC) [10]	1	80
	2	123
	3	44
	4	22
pN (AJCC)[10]	0	88
	1	170
	2	11
Nodes positive (Pathological)	0	72
	1-3	79
	4-9	43
	≥10	75
Response to treatment	Responders	10
	None	23
5-year disease-free survival	Yes	191
	No	44
	Censored	1
	(remaining are in stage IV)	(33)

plasmic % Intensity defines percent intensity of stain within cytoplasm; Cytoplasmic % Positive defines percent of all cells positive within cytoplasm; Cytoplasmic Total Score defines the product of Cytoplasmic % Intensity and Cytoplasmic % Positive; Nuclear % Intensity defines percent intensity of stain within nucleus; Nuclear % Positive defines percent of all cells positive within nucleus; and Nuclear Total Score defines the product of Nuclear % Intensity and Nuclear % Positive.

Anti-EGFR antibody, anti-HER2/neu antibody, anti-phospho-EGFR antibody (Tyr845), and anti-phospho-HER2/neu (Tyr877) were from Cell Signaling Technology, Inc. (Beverly, MA).

Machine learning algorithms

Feature selection algorithms, *Gain Ratio Attribute Evaluation* and *Relief* implemented in WEKA 3.4 (3) (http://www.cs.waikato.ac.nz/ml/weka/) were used to identify important signal protein subsets and clinical factors. The *Random Committee* algorithm implemented in software package WEKA 3.4 was used to construct the individualized survival and treatment response prediction models. The random committee algorithm builds an ensemble of random classification trees and averages their predictions. Ten-fold cross-validation was used to evaluate the performance of the prediction models. The details of the bioinformatics analysis are provided in the Supplementary Information.

Statistical methods

To assess the significance of individual classifier performance, we computed the probability of the observed prediction accuracy occurring by chance (random prediction using a fair coin flip). The probability of doing at least as well as our prediction models by chance was calculated using *Binomial Distribution* functions in software package *R* (http://www.r-project.org/). Statistical significance test was used to evaluate different prediction results on the same cohort.

RESULTS

Identifying important clinical factors and activated protein kinase expression profiles in breast cancer survival

Breast cancer patients with the same stage of disease can vary markedly in respect of the chance of developing metastatic and recurrent disease after surgery. High-risk patients should be given closer follow-up checks and more aggressive treatments such as adjuvant chemotherapy. In order to develop a powerful profile to accurately classify patients into subgroups of good prognosis and poor prognosis, we first ranked the importance of the

clinical parameters in breast cancer survival using the *Gain Ratio* attribute evaluation algorithm. The optimal subset of clinical factors was identified by sequentially adding the top ranked parameters to the survival prediction model until the highest prediction accuracy was reached. Our results showed that the optimal subset of clinical factors for breast cancer survival prediction include histology, positive lymph node status, pT (AJCC), pN (AJCC) (10), and smoking (the sequence represents the order of the ranking) (Tab. II).

To identify the best subset of activated protein kinases in predicting breast cancer survival, 42 antibody scores representing the expression levels of 7 activated protein kinases were ranked by using the *Gain Ratio* algorithm. The results showed that the top 7 antibody scores were the most informative predictors for breast cancer survival. These 7 antibody scores contained the measurements for the level of 6 activated protein kinases: phospho-EGFR, phospho-ER, phospho-HER2/neu, phospho-IGF-IR/In, phospho-MAPK, and phospho-p70^{S6K} (Tab. II).

Disease-free survival prediction of individual breast cancer patients

Accurate prediction of a patient's risk of developing metastatic or recurrent disease aids individualized decision-making for prescribing expensive and toxic adjuvant chemotherapy to those at high risk and avoiding overtreatment of those at low risk. To identify the optimal classifier for individualized survival prediction in patients with breast cancer, we performed the following analyses using the *Random Committee* algorithm: (1) as control, we used the expression profiles of all 7 activated protein kinases with a total of 42 antibody scores to predict whether or not the patient would survive a 5-year interval; (2) in addition to the expression profiles of the 7 activated protein kinases measured by the 42 antibody scores, the selected clinical factors (Tab. II) were added to the survival prediction model; and (3) instead of using all 7 activated protein kinases, only the top ranked 7 antibody scores for 6 activated protein kinases plus the identified clinical factors were used as predictors for the individualized survival prediction (Tab. II).

Our results showed that, when the expression profiles of all 7 activated protein kinases were used, the overall accuracy of survival prediction was 82.5%, with a sensitivity (prediction accuracy of 5-year survival or low risk) of 98.5% and a specificity (prediction accuracy of non-5-year survival or high risk) of 9.1% (Tab. III, Analysis 1). When the selected clinical-pathological markers were added to the prediction model, the overall accuracy increased to 85.8% ($p<0.16$), with a sensitivity of 99% and a specificity of 25% (Tab. III, Analysis 2). Furthermore, our results indicated that, when the identified 7 antibody scores for the 6 activated protein kinases plus the selected clinical parameters were used (Tab. II), the overall accuracy of survival prediction increased significantly from 82.5% to 92.3% ($p<0.0005$) and the specificity increased from 9.1% to 65.9% ($p<1E-7$) (Tab. III, Analysis 3). Our study demonstrated that, with the use of the identified expression profiles of phospho-EGFR, phospho-ER, phospho-HER2/neu, phospho-IGF-IR/In, phospho-MAPK, and phospho-p70^{S6K} as well as the selected clinical information we were able to accurately classify breast cancer patients into subgroups of good prognosis and poor prognosis.

TABLE II - TOP RANKED CLINICAL FACTORS AND ACTIVATED PROTEIN KINASE EXPRESSION PROFILES IN BREAST CANCER SURVIVAL PREDICTION

Clinical information	Histology Nodes positive pT pN Smoking
Protein expression measurements	phospho-IGF-IR/In Nuclear % Intensity phospho-p70^{S6K} Cytoplasmic % Intensity phospho-p70^{S6K} Cytoplasmic Total Score phospho-ER Cytoplasmic % Intensity phospho-HER2/neu Cytoplasmic % Positive phospho-MAPK Nuclear % Intensity phospho-EGFR Cytoplasmic % Positive

Identifying important clinical factors and activated protein kinase expression profiles for individualized treatment response prediction

Precise prediction of a patient's response to certain therapeutic regimens is crucial to devise the most appropriate treatment combination for each individual breast cancer patient. To achieve the goal of personalized medicine, it is vital to identify important clinical-pathological parameters and novel molecular targets that are predictive of a patient's response to treatments. The aim of this study is to develop a prediction model of treatment response based on existing data. With the establishment of this model, physicians can assess a patient's predisposition to response to a certain chemotherapeutic agent in a prospective clinical trial. Specifically, based on the patient's clinical-pathological traits and molecular signature, a treatment option can be input to the model as a predictor, and the output "predicted response" will give an indication of the patient's predisposition to response to this agent. By substituting all possible treatment options into the prediction model (1 at a time), the physicians will be able to identify the effective treatment strategy. From the machine-learning aspect, the knowledge base should incorporate different pharmacological treatments and the entailed responses of the specific patients in retrospective studies, such that the predictive model

TABLE III - SURVIVAL PREDICTION FOR INDIVIDUAL BREAST CANCER PATIENTS USING ACTIVATED PROTEIN KINASE EXPRESSION PROFILES AND CLINICAL-PATHOLOGICAL PARAMETERS (N=235)

		Analysis 1	Analysis 2	Analysis 3
Predictors	phospho-Akt	x	x	
	phospho-p70^{S6K}	x	x	x
	phospho-ER	x	x	x
	phospho-EGFR	x	x	x
	phospho-IGF-IR/In	x	x	x
	phospho- HER2/neu	x	x	x
	phospho-MAPK	x	x	x
	Clinical information		x	x
Prediction accuracy	Sensitivity (5-year survival)	98.5%	99.0%	98.0%
	Specificity (non-5-year survival)	9.1%	25.0%	65.9%
	Overall accuracy	82.5%	85.8%	92.3%
Significance of overall accuracy		p<5.1E-27	p<6.4E-33	p<8.6E-48

Analysis 1: using all the antibody scores of the 7 activated signaling proteins to predict disease-free survival. Analysis 2: using all the antibody scores and the selected clinical factors in Table II to predict disease-free survival. Analysis 3: using the selected antibody scores (Tab. II) and the selected clinical factors (Tab. II) to predict disease-free survival.

can identify the effective treatment options for a given patient in a prospective analysis.

In order to identify the most important clinical-pathological information for individualized treatment response prediction, we applied the *Gain Ratio* attribute evaluation algorithm to rank the importance of the clinical parameters in the studied cohort. The results showed that the top 9 clinical factors are the most informative in treatment response prediction, including metastasis site, smoking, ER, histology, PR, surgery procedure, chemotherapy, stage, and pN (the sequence represents the order of the ranking) (Tab. IV).

To identify the optimal subset of activated protein kinases in treatment response prediction, we first ranked the importance of the 42 antibody scores for the 7 activated protein kinases using *Gain Ratio*. The results showed that only 4 antibody scores were informative in treatment response prediction. These included the measurement for phospho-ER, phospho-EGFR, phospho-HER2/neu, and phospho-p70^{S6K}. These 4 protein kinases were chosen as predictors for treatment response. We then used the *Relief* algorithm to rank the 42 antibody scores in treatment response prediction. The top 7 antibody scores ranked by *Relief* were consistent with those using *Gain Ratio* with regard to the protein kinases. We then added the top antibody scores ranked by *Relief* to the treatment response prediction model one by one. The results showed that when the antibody score *phospho-MAPK Cytoplasmic % Positive* was added to the prediction model, the subset of the protein kinases, containing phospho-ER, phospho-EGFR, phospho-HER2/neu, phospho-MAPK and phospho-p70^{S6K} (Tab. IV), generated optimal treatment response prediction in supervised training.

TABLE IV - TOP RANKED CLINICAL FACTORS AND ACTIVATED PROTEIN KINASE EXPRESSION PROFILES IN BREAST CANCER TREATMENT RESPONSE PREDICTION

Clinical information	Metastasis site
	Smoking
	ER
	Histology
	PR
	Surgery procedure
	Chemotherapy
	Stage
	pN
Protein expression measurements	phospho-ER Nuclear % Positive
	phospho-HER2/neu Nuclear % Positive
	phospho-p70^{S6K} Cytoplasmic % Positive
	phospho-EGFR Cytoplasmic % Positive
	phospho-MAPK Cytoplasmic % Positive

Treatment response prediction of individual breast cancer patients

To construct the optimal classifier for treatment response prediction for individual breast cancer patients, we performed the following analyses using *Random Committee*: (1) we used the expression profiles of all 7 activated protein kinases (with 42 antibody scores) to predict a patient's response to treatments (i.e., responder or nonresponder); (2) in addition to the protein expression profiles, the identified clinical factors (Tab. IV) were included in the prediction model to assess a patient's response to treatments; and (3) instead of using the com-

TABLE V - TREATMENT RESPONSE PREDICTION FOR INDIVIDUAL BREAST CANCER PATIENTS USING ACTIVATED PROTEIN KINASE EXPRESSION PROFILES AND CLINICAL-PATHOLOGICAL PARAMETERS (N=33)

		Analysis 1	Analysis 2	Analysis 3
Predictors	phospho-Akt	x	x	
	phospho-p70^{S6K}	x	x	x
	phospho-ER	x	x	x
	phospho-EGFR	x	x	x
	phospho-IGF-IR/In	x	x	
	phospho-HER2/neu	x	x	x
	phospho-MAPK	x	x	x
	Clinical information		x	x
Prediction accuracy	Sensitivity (responder)	23.9%	33.8%	78.9%
	Specificity (non-responder)	98.0%	99.5%	97.5%
	Overall accuracy	78.5%	82.2%	92.6%
Significance of overall accuracy		p<1.7E-23	p<1.0E-29	p<1.0E-54

Analysis 1: using all the antibody scores of the 7 activated signaling proteins to predict treatment response. Analysis 2: using all the antibody scores and the selected clinical factors in Table IV to predict treatment response. Analysis III: using the selected antibody scores (Tab. IV) and the selected clinical factors (Tab. IV) to predict treatment response.

plete protein expression profiles, the top ranked protein expression profiles (Tab. IV) for 5 activated protein kinases plus the selected clinical information were used as predictors to assess a patient's response to treatments.

Our results showed that when the complete expression profiles for the 7 activated protein kinases were used, the treatment response prediction accuracy was 78.5%, with sensitivity 23.9% and specificity 98% (Tab. V, Analysis 1). When the selected clinical information was used together with the expression profiles of the 7 activated protein kinases, the treatment response prediction accuracy increased to 82.2% ($p<0.35$), with sensitivity 33.8% and specificity 99.5% (Tab. V, Analysis 2). Using the identified protein markers plus the selected clinical parameters, the overall accuracy of treatment response prediction significantly increased from 78.9% to 92.6% ($p<0.05$) and the sensitivity increased from 23.9% to 78.9% ($p<0.007$) (Tab. V, Analysis 3). Our results demonstrated that, using the identified signal protein expression profiles of phospho-EGFR, phospho-ER, phospho-HER2/neu, phospho-MAPK and phospho-p70^{S6K} along with the selected clinical information, we were able to precisely predict a patient's response to certain treatment options.

DISCUSSION

Accurate prediction of clinical outcome depends on suitable computational models. To construct reliable prediction models, the first important task is to identify informative and relevant predictors. Good feature selection algorithms can identify important predictors while taking into account the interactions among them. In this study, 2 feature selection algorithms, *Relief* and *Gain Ratio* attribute selection, were used to evaluate the importance of several activated protein kinases as well as patients' clinical information in breast cancer outcome predictions. Our results demonstrated that the identified subsets of the activated protein kinases significantly ($p<0.0005$) increased the accuracy of clinical outcome predictions. In addition, the *Random Committee* algorithm predicted each individual patient's survival and treatment responses with overall accuracy above 92% for the study cohort. Together, our model system provided a comprehensive profile encompassing multiple signal pathways and several important clinical-pathological parameters, and accurately predicted the outcome of individual breast cancer patients.

In this study, we found that both EGFR (ErbB1/HER1) and HER2/neu (ErbB2) were among the important parameters of survival and treatment response predictions for breast cancers. Both proteins belong to the epidermal growth factor receptor (EGFR) family (11). There is an increasing body of evidence showing that the EGFR family plays an essential role in breast cancer development and progression (12). It was found that EGFR overexpression was present in 14-90% of breast cancers and was linked to poor prognosis, depending on different samples and the methods by which receptors are quantified. HER2/neu overexpression was correlated with adverse prognosis in both node-negative and node-positive breast cancers (13). In vivo and in vitro studies demonstrated that overexpression of HER2/neu in transgenic mouse mammary glands promoted oncogenic transformation and development of a malignant phenotype, and overexpression of HER2/neu in breast cancer cell lines increased the metastatic and invasive potential (12). There-

fore, our selection of both EGFR and HER2/neu as parameters of breast cancer outcome prediction is relevant to the pathogenic roles of these 2 proteins in the development and progression of breast cancer.

In addition to EGFR and HER2/neu, we also found that both mitogen-activated protein kinase (MAPK) and $p70^{S6K}$ were important protein kinases in breast cancer survival and treatment response predictions. MAPK is one of the major downstream signaling proteins of the EGFR family. It was demonstrated that the enhancement of EGFR/HER2/neu heterodimerization and receptor activation was correlated with increased activation of MAPK in TAMR cells. Furthermore, the activation of MAPK was associated with poor response to antihormonal therapy and decreased survival in breast cancers (14). Another major downstream signaling pathway of the EGFR family is PI3K-Akt-$p70^{S6K}$. The activation of PI3K-Akt-$P70^{S6K}$ was critical for EGFR-induced proliferative response in breast cancer cell lines, and the inhibition of this pathway with rapamycin substantially inhibited cell cycle progression, cell growth, and cell proliferation in breast cancer cell lines (15).

Bidirectional cross-talk between EGFR receptor pathways and ER signaling pathways in breast cancers is well established. Studies found that EGFR receptors had an ability to enhance ER signaling either by direct activation or through MAPK activation, while ER had an ability to mediate EGFR-induced MAPK activation through the regulation of TGFa availability (16, 17). The simultaneous selection of EGFR receptor family and ER in this study proved the importance of the cross-talk between these 2 groups of proteins in breast cancer development and progression and demonstrated the relevance of our selected biomarkers in the pathogenesis of breast cancers. In this study, IGF-IR/In was also selected in risk assessment for breast cancer. Expression of IGF-IR/In was found in the majority of breast cancers. It was shown that IGF-IR/In was one of the most potent mitogens to breast cancer cells in vitro (18). Furthermore, it was found that IGF-IR/In and ER reciprocally engaged in a powerful functional cross-talk to enhance signal transduction in breast cancer development and progression (19).

Most notably in this study, we evaluated protein phosphorylation levels instead of total protein expression levels. All proteins selected in this study belong to protein kinases. Protein phosphorylation and dephosphorylation are well-characterized biochemical processes for protein kinases to conduct cellular signal transduction. With regard to the EGFR receptor family, the autophosphorylation in tyrosines is an essential step for their activation and is a strict requirement for them to transform cells. The level of auto-tyrosine phosphorylation directly correlates with the scale of the activation of the EGFR receptor family. Likewise, phosphorylation at certain tyrosine, serine or threonine residues in other kinases is a key step for their activation and the measurements of these phosphorylations reflect their functional status in vivo. Thus, the detection of protein kinase phosphorylation levels more accurately reveals the molecular mechanisms of breast cancer, which is more significant for individualized survival and treatment response prediction and potentially provides the clues at the molecular level for the selection of optimal therapy of breast cancers.

ACKNOWLEDGMENTS

We would like to thank Mr Steve Turner, the former CEO of Clinomics Biosciences Inc., for supporting this project. We thank Dr Patrick Muraca, the former COO of Clinomics, for his thoughtful discussions. Brad Vincent compiled the data file from the Cell Signaling Database maintained by the Clinomics Biosciences Inc.

This project is supported by the NIH/NCRR P20 RR16440-03 grant.

DISCLAIMER

The findings and conclusions in this report are those of the author(s) and do not necessarily represent the views of the National Institute for Occupational Safety and Health.

Novelty and impact: This paper provides a novel bioinformatic scheme which integrates several state-of-the-art machine learning algorithms for early detection of high-risk breast cancers and accurate prediction of treatment response. The prediction system combines multiple signal protein activation profiles and relevant clinical information, and provides a unique guideline to aid individualized decision-making in the clinical management of breast cancer.

Address for correspondence:
For bioinformatics contact
Lan Guo
MBR Cancer Center/Department of Community Medicine
West Virginia University
Morgantown, WV 26505, USA
e-mail: lguo@hsc.wvu.edu

For cancer biology contact
Yong Qian
The Pathology and Physiology Research Branch
Health Effects Laboratory Division
National Institute for Occupational Safety and Health
Morgantown, WV 26505, USA
e-mail: yaq2@cdc.gov

REFERENCES

1. Murphy N, Millar E, Lee CS. Gene expression profiling in breast cancer: towards individualising patient management. Pathology 2005; 37: 271-7.
2. Rosen JM. Hormone receptor patterning plays a critical role in normal lobuloalveolar development and breast cancer progression. Breast Dis 2003; 18: 3-9.
3. Witten IH, Frank E. Data mining: Practical machine learning tools and techniques (2nd edition). San Francisco: Morgan Kaufmann 2005.
4. Warford A. Tissue microarrays: fast-tracking protein expression at the cellular level. Expert Rev Proteomics 2004; 1: 283-92.
5. Warford A, Howat W, McCafferty J. Expression profiling by high-throughput immunohistochemistry. J Immunol Methods 2004; 290: 81-92.
6. Therasse P, Arbuck SG, Eisenhauer EA, et al. New guidelines to evaluate the response to treatment in solid tumors. European Organization for Research and Treatment of Cancer, National Cancer Institute of the United States, National Cancer Institute of Canada. J Natl Cancer Inst 2000; 92: 205-16.
7. Kallakury BV, Sheehan CE, Ambros RA, et al. Correlation of p34cdc2 cyclin-dependent kinase overexpression, CD44s downregulation, and HER-2/neu oncogene amplification with recurrence in prostatic adenocarcinomas. J Clin Oncol 1998; 16: 1302-9.
8. Ouban A, Muraca P, Yeatman T, Coppola D. Expression and distribution of insulin-like growth factor-1 receptor in human carcinomas. Hum Pathol 2003; 34: 803-8.
9. Allred DC, Clark GM, Elledge R, et al. Association of p53 protein expression with tumor cell proliferation rate and clinical outcome in node-negative breast cancer. J Natl Cancer Inst 1993; 85: 200-6.
10. American Joint Committee on Cancer Staging Manual, 5th ed. Philadelphia, PA: Lippincott-Raven 1997.
11. Zaczek A, Brandt B, Bielawski KP. The diverse signaling network of EGFR, HER2, HER3 and HER4 tyrosine kinase receptors and the consequences for therapeutic approaches. Histol Histopathol 2005; 20: 1005-15.
12. Atalay G, Cardoso F, Awada A, Piccart MJ. Novel therapeutic strategies targeting the epidermal growth factor receptor (EGFR) family and its downstream effectors in breast cancer. Ann Oncol 2003; 14: 1346-63.
13. Ross JS, Linette GP, Stec J, et al. Breast cancer biomarkers and molecular medicine. Expert Rev Mol Diagn 2003; 3: 573-85.
14. Gee JM, Robertson JF, Ellis IO, Nicholson RI. Phosphorylation of ERK1/2 mitogen-activated protein kinase is associated with poor response to anti-hormonal therapy and decreased patient survival in clinical breast cancer. Int J Cancer 2001; 95: 247-54.
15. Mita MM, Mita A, Rowinsky EK. The molecular target of rapamycin (mTOR) as a therapeutic target against cancer. Cancer Biol Ther 2003; 2 (4 Suppl 1): S169-77.
16. Bunone G, Briand PA, Miksicek RJ, Picard D. Activation of the unliganded estrogen receptor by EGF involves the MAP kinase pathway and direct phosphorylation. EMBO J 1996; 15: 2174-83.
17. Hutcheson IR, Knowlden JM, Madden TA, et al. Oestrogen receptor-mediated modulation of the EGFR/MAPK pathway in tamoxifen-resistant MCF-7 cells. Breast Cancer Res Treat 2003; 81: 81-93.
18. Lonning PE, Helle SI. IGF-1 and breast cancer. Novartis Found Symp 2004; 262: 205-12.
19. Surmacz E, Bartucci M. Role of estrogen receptor alpha in modulating IGF-I receptor signaling and function in breast cancer. J Exp Clin Cancer Res 2004; 23: 385-94.

APPENDIX

MACHINE LEARNING CLASSIFIER

We used the *Random Committee* in WEKA3.4 (1) to construct the classifiers in this study. The *Random Committee* algorithm is an ensemble of random classification trees. Each tree is based on the same training data but uses a different random number seed to build the base classifier. Each classifier generates a probability estimate, which is averaged as the final prediction result. This model is based on *bagging* and randomization, which introduces variability into each single base classifier. When the base classification trees are combined in the *Random Committee* algorithm, the output is generally more accurate than a single classification tree (1). The prediction accuracy was evaluated using 10-fold cross-validation. In the 10-fold cross-validation, the data set was randomly partitioned into 10 folds of equal size with possible exception of the last fold (the last fold contains the remaining samples). The prediction models were trained and tested 10 times. Each time, 9 folds were picked to build the prediction model, while the remaining fold was validated on the prediction model. We used 10-fold cross-validation to evaluate the prediction models in this study, because the estimation accuracy by this validation method has been proven to have the lowest bias and variance among all validation methods, including the leave-one-out method (2). It thus provides an objective evaluation of the performance of our prediction models in general.

FEATURE SELECTION ALGORITHMS

In this study, we used the *Gain Ratio* attribute selection algorithm and the *Relief* algorithm in WEKA3.4 to rank the importance of clinical information and antibody scores in breast cancer outcome predictions.

Relief is an instance-based attribute ranking scheme. It samples an instance randomly from the data and checks its nearest neighbor from the same and opposite class. The values of the attributes of the nearest neighbors are compared to the sampled instance and used to update the relevance scores for each attribute. This process is repeated for a user-specified number of *m* times (for *m* number of instances). The rationale is that an informative attribute should have the same value for instances from the same class and differentiate between instances from different classes (1, 3).

The *Gain Ratio* attribute selection algorithm ranks the importance of individual attributes in the classification. It was originally used with decision-tree classification (4). Suppose the training set contains *p* and *n* objects of class *P* and *N*, respectively. Let attribute *A* have values $A_1, A_2,\ldots A_v$ and let the number of objects with value A_i of attribute *A* be p_i and n_i (corresponding to class *P* and *N*), respectively. The value of attribute *A* can be expressed as Equation 1:

$$IV(A) = -\sum_{i=1}^{v} \frac{p_i + n_i}{p+n} \log_2 \frac{p_i + n_i}{p+n} \qquad \textit{(Equation 1)}$$

Another criterion, *Gain(A),* measures the reduction in the information requirement for a classification rule if the decision tree uses attribute *A* as a root. The information required to make a classification by attribute *A* is measure by Equation 2:

$$I(p,n) = -\frac{p}{p+n}\log_2\frac{p}{p+n} - \frac{n}{p+n}\log_2\frac{n}{p+n} \qquad \textit{(Equation 2)}$$

The expected information required for the tree with *A* as root is then obtained as the weighted average as in Equation 3:

$$E(A) = \sum_{i=1}^{v} \frac{p_i + n_i}{p+n} I(p_i, n_i) \qquad \textit{(Equation 3)}$$

The information gained by branching on *A* is therefore:

$$Gain(A) = I(p,n) - E(A) \qquad \textit{(Equation 4)}$$

The importance of variable *A* is measured by the ratio:

$$Gain(A)/IV(A) \qquad \textit{(Equation 5)}$$

the larger the value the more important variable A is.

IMPORTANT CLINICAL INFORMATION AND PROTEIN EXPRESSION PROFILES

Identifying important profiles for disease-free survival prediction

Using the *Gain Ratio* attribute selection algorithm implemented in WEKA3.4, we ranked the clinical parameters in survival prediction (Tab. I). The search method was *ranker*. The clinical parameters were added to the prediction model in a stepwise manner (first add top 1, and then top 2, top 3, etc.). The top 5 parameters (bold, Tab. I) resulted in optimal prediction results and thus were selected for disease-free survival prediction of individual breast cancer patients.

Forty-two antibody measurement scores were also ranked by using the *Gain Ratio* attribute selection algorithm implemented in WEKA3.4. The search method was *ranker*. The 10-fold cross-validation method was used to evaluate the average ranking of each antibody score (Tab. II). The top 7 antibody scores were selected as survival predictors of individual breast cancer patients, because they were the most informative in survival prediction and resulted in the highest accuracy.

Identifying important profiles for treatment response prediction

Using the *Gain Ratio* attribute selection algorithm implemented in WEKA3.4, we ranked the clinical parameters in treatment response prediction (Tab. III). The search method was *ranker*. The top 9 parameters (with *Gain Ratio* value greater than 0) were the most informative in treatment response prediction and resulted in highest accuracy. They were thus selected for treatment response prediction of individual breast cancer patients.

Forty-two antibody measurement scores were also ranked by using the *Gain Ratio* attribute selection algorithm and the *Relief* algorithm implemented in WEKA3.4. Both attribute evaluators used the search method *ranker*. Only 4 antibody scores had merit greater than 0 by using

TABLE I - RANKING OF THE CLINICAL PARAMETERS IN SURVIVAL PREDICTION

Clinical parameters	Merit (*Gain Ratio* value)
Histology	**0.105809**
Nodes positive	**0.05924**
pT	**0.04056**
pN	**0.028842**
Smoking	**0.027926**
Stage	0.02325
PR	0.002837
ER	0.000301
Age	0
HER2/neu	0

TABLE II - RANKING OF THE 42 ANTIBODY SCORES IN SURVIVAL PREDICTION

average merit	average rank	attribute
0.03 +- 0.002	3.3 +- 1.55	phospho-IGF-IR/In Nuc Intensity
0.028 +- 0.018	7 +- 8.53	phospho-p70S6K Cyto Total
0.015 +- 0.003	7.4 +- 1.8	phospho-Her2/neu Cyto Positive
0.018 +- 0.006	8.1 +- 5.11	phospho-MAPK Nuc Intensity
0.021 +- 0.007	8.5 +- 9	phospho-ER Cyto Intensity
0.023 +- 0.015	9.4 +- 8.97	phospho-p70S6K Cyto Intensity
0.001 +- 0.004	11.2 +- 3.79	phospho-EGFR Cyto Positive
0 +- 0	12.4 +- 5.28	phospho-p70S6K Nuc Total
0 +- 0	13.2 +- 5.21	phospho-p70S6K Nuc Intensity
0 +- 0	13.4 +- 1.43	phospho-EGFR Nuc Intensity
0 +- 0	13.5 +- 2.94	phospho-EGFR Cyto Intensity
0 +- 0	14.3 +- 1.19	phospho-EGFR Cyto Total
0.017 +- 0.011	14.6 +-13.46	phospho-ER Cyto Total
0.023 +- 0.019	16.4 +-18.49	phospho-Her2/neu Cyto Total
0 +- 0	16.6 +- 9.99	phospho-EGFR Nuc Intensity
0 +- 0	16.8 +- 1.94	phospho-Akt Nuc Positive
0 +- 0	18.2 +- 1.94	phospho-Akt Cyto Total
0.02 +- 0.017	18.3 +-18.34	phospho-Her2/neu Cyto Intensity
0 +- 0	18.4 +- 1.74	phospho-Akt Cyto Positive
0 +- 0	18.8 +- 4.49	phospho-p70S6K Nuc Positive
0 +- 0	19.4 +- 3.38	phospho-Akt Nuc Intensity
0 +- 0	20.1 +- 1.92	phospho-Akt Cyto Intensity
0.007 +- 0.01	20.8 +-10.36	phospho-MAPK Nuc Total
0.003 +- 0.004	22.3 +-11.38	phospho-MAPK Cyto Positive
0.009 +- 0.014	22.5 +-11.49	phospho-IGF-IR/In Nuc Total
0 +- 0	22.9 +- 1.92	phospho-Akt Nuc Total
0 +- 0	23.8 +- 0.75	phospho-p70S6K Cyto Posive
0 +- 0	26.3 +-12.53	phospho-EGFR Nuc Total
0 +- 0	28.1 +- 4.41	phospho-IGF-IR/In Cyto Positive
0 +- 0	28.3 +- 2.33	phospho-IGF-IR/In Cyto Intensity
0 +- 0	29.9 +- 3.05	phospho-IGF-IR/In Nuc Positive
0 +- 0	30.1 +- 3.01	phospho-IGF-IR/In Cyto Total
0 +- 0	31.2 +- 5.21	phospho-Her2/neu Nuc Intensity
0 +- 0	31.4 +- 5.1	phospho-Her2/neu Nuc Total
0 +- 0	33.1 +- 3.01	phospho-MAPK Cyto Intensity
0 +- 0	33.5 +- 7.89	phospho-ER Cyto Positive
0 +- 0	33.9 +- 3.91	phospho-MAPK Nuc Positive
0 +- 0	35.5 +- 3.2	phospho-MAPK Cyto Total
0 +- 0	36.3 +- 5.75	phospho-Her2/neu Nuc Positive
0 +- 0.001	36.5 +- 9.12	phospho-ER Nuc Intensity
0 +- 0	37 +- 2.9	phospho-ER Nuc Positive
0 +- 0	40.3 +- 2.45	phospho-ER Nuc Total

TABLE III - RANKING OF CLINICAL PARAMETERS IN TREATMENT RESPONSE PREDICTION USING *GAIN RATIO* ALGORITHM

Clinical parameters	Merit (*Gain Ratio* value)
Metastasis site	**0.349254**
Smoking	**0.144443**
ER	**0.127986**
Histology	**0.069955**
PR	**0.069406**
Surgery procedure	**0.037344**
Chemotherapy	**0.01126**
Stage	**0.010424**
pN	**0.009036**
Age	0
pT	0
Radiation	0
HER2/neu	0
Nodes positive	0

TABLE IV - TOP RANKED ANTIBODY SCORES IN TREATMENT RESPONSE PREDICTION USING *GAIN RATIO* ALGORITHM

Antibody score	Merit (*Gain Ratio* value)
Phospho-ER Nuclear % Intensity	0.0585
Phospho-HER2/neu Nuclear % Positive	0.047
Phospho-p70^{S6K} Cytoplasmic % Positive	0.0384
Phospho-EGFR Cytoplasmic % Positive	0.0283

The remaining antibody scores all had merit equal to 0.

the Gain Ratio algorithm (Tab. IV). They were selected as predictors of a patient's response to treatment. Top ranked antibody scores using *Relief* (Tab. V) were also added to the prediction model in a stepwise manner. The top scores were first added one-by-one, separately. Then

TABLE V - TOP RANKED ANTIBODY SCORES IN TREATMENT RESPONSE PREDICTION USING *RELIEF* ALGORITHM

Antibody score	Merit
phospho-ER Cytoplasmic % Positive	0.064286
phospho-HER2/neu Cytoplasmic % Total	0.053016
phospho-p70^{S6K} Cytoplasmic % Total	0.049563
phospho-ER Cytoplasmic % Total	0.047455
phospho-HER2/neu Cytoplasmic % Intensity	0.045476
phospho-p70^{S6K} Cytoplasmic % Intensity	0.042143
phospho-p70^{S6K} Cytoplasmic % Positive	0.031786
phospho-MAPK Cytoplasmic % Positive	0.031667
phospho-MAPK Cytoplasmic % Intensity	0.030556
phospho-EGFR Cytoplasmic Total	0.029643
phospho-ER Nuclear % Intensity	0.0275

the top scores were added incrementally: first top 1, then top 2, top 3, etc. Our results showed that when pMAPK Cytoplasmic % Positive was added to the prediction model, this subset of antibody scores had the highest accuracy in treatment response prediction. Therefore, the final subset of antibody scores for treatment response prediction was comprised of phospho-ER Nuclear % Intensity, phospho-HER2/neu Nuclear % Positive, phospho-p70^{S6K} Cytoplasmic % Positive, phospho-EGFR Cytoplasmic % Positive, and phospho-MAPK Cytoplasmic % Positive.

REFERENCES

1. Witten IH, Frank E. Data mining: Practical machine learning tools and techniques (2nd edition). San Francisco: Morgan Kaufmann 2005.
2. Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection. Proceedings of International Joint Conference on Artificial Intelligence (IJCAI) 1995; 1137-43.
3. Hall MA, Holmes G. Benchmarking attribute selection techniques for discrete class data mining. IEEE Transactions on Knowledge and Data Engineering 2003; 15: 1437-47.
4. Quinlan JR. Induction of decision tree. Machine Learning 1986; 1: 81-106.

Received: October 20, 2006
Accepted: January 6, 2007