Predicting continuous values of prognostic markers in breast cancer from microarray gene expression profiles

Sofia K. Gruvberger-Saal,[1] Patrik Edén,[2] Markus Ringnér,[2,3] Bo Baldetorp,[1] Gunilla Chebil,[4] Åke Borg,[1] Mårten Fernö,[1] Carsten Peterson,[2] and Paul S. Meltzer[3]

[1]Department of Oncology and [2]Complex Systems Division, Department of Theoretical Physics, Lund University, Lund, Sweden; [3]Cancer Genetics Branch, National Human Genome Research Institute, NIH, Bethesda, MD; and [4]Department of Pathology, Helsingborg Hospital, Helsingborg, Sweden

Abstract

The prognostic and treatment-predictive markers currently in use for breast cancer are commonly based on the protein levels of individual genes (*e.g.*, steroid receptors) or aspects of the tumor phenotype, such as histological grade and percentage of cells in the DNA synthesis phase of the cell cycle. Microarrays have previously been used to classify binary classes in breast cancer such as estrogen receptor (ER)-α status. To test whether the properties and specific values of conventional prognostic markers are encoded within tumor gene expression profiles, we have analyzed 48 well-characterized primary tumors from lymph node-negative breast cancer patients using 6728-element cDNA microarrays. In the present study, we used artificial neural networks trained with tumor gene expression data to predict the ER protein values on a continuous scale. Furthermore, we determined a gene expression profile-directed threshold for ER protein level to redefine the cutoff between ER-positive and ER-negative classes that may be more biologically relevant. With a similar approach, we studied the prediction of other prognostic parameters such as percentage cells in the S phase of the cell cycle (SPF), histological grade, DNA ploidy status, and progesterone receptor status. Interestingly, there was a consistent reciprocal relationship in expression levels of the genes important for both ER and SPF prediction. This and similar studies may be used to increase our understanding of the biology underlying these markers as well as to improve the currently available prognostic markers for breast cancer. [Mol Cancer Ther. 2004;3(2):161–168]

Introduction

For prognostic evaluation of breast cancer, the most important information is given by the size of the primary tumor, involvement of regional lymph nodes, and occurrence of distant metastasis. Histopathological and biochemical markers constitute important tools for identifying the group of aggressive breast cancers with a poor prognosis and for predicting the response to treatment. Treatment-predictive and prognostic markers currently in use for breast cancer include variables such as histological grade, age, steroid receptors, and markers of proliferation represented by the fraction of cells in the S phase of the cell cycle (SPF) or thymidine labeling index. However, the genes and pathways associated with these markers are not sufficiently known and the processes that lead to their clinical manifestation are not fully understood.

Multiparametric methods such as microarray analysis, which rely on many pieces of information, seem ideally suited for grouping of tumor subtypes. Indeed, the microarray technique has successfully been used to classify breast cancer into different subgroups with clinical correlations (1–3) as well as using the expression profiles to predict cancer types and disease recurrence of patients (4–7). In general, these studies use statistical methods to generate an output, which classifies a sample as a member of one group or another. Expression profiles have thus far not been used to provide a graded output corresponding to the continuum of biological properties exhibited by tumors.

Although prognostic markers for breast tumors are used to categorize tumors into two groups [*e.g.*, estrogen receptor (ER) positive *versus* ER negative or high SPF *versus* low SPF], in reality, these subdivisions are defined by applying cutoff values to a continuous laboratory value. For example, the cutoff values used to subgroup tumors based on ER status are defined from clinical studies correlating ER values with response to endocrine treatment and are not based on measurements of the functional activity of the ER signal transduction pathway. In this study, we have investigated the possibility of predicting not only the binary ER status and SPF of a tumor but also the continuous values of ER protein and SPF from gene expression profiles. We have used cDNA microarrays and artificial neural networks (ANNs) to analyze the expression of 6728 genes in 48 well-characterized primary tumors representing a broad spectrum of ER protein expression and SPF values. From the results of these predictions, we have generated ranked lists of the genes most sensitive for the predictions and defined a cutoff for ER status based on gene expression. Furthermore, using a similar approach, we have studied the gene

Received 6/24/03; revised 11/3/03; accepted 11/4/03.

Grant support: Supported in part by the Lund University Medical Faculty, the Swedish Cancer Society, the Berta Kamprads Foundation, the Gunnar Arvid & Elisabeth Nilsson Foundation, the Hospital of Lund Foundations, the E and F Bergqvist Foundation, the King Gustav V's Jubilee Foundation, the Swedish Foundation for Strategic Research, the Swedish Research Council, and the Knut and Alice Wallenberg Foundation through the SWEGENE consortium.

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked advertisement in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Note: S. K. Gruvberger-Saal and P. Edén contributed equally.

Requests for Reprints: Paul S. Meltzer, Section of Molecular Genetics, Cancer Genetics Branch, National Human Genome Research Institute, NIH, MSC 8000, Room 5139, 50 South Drive, Bethesda, MD 20892-8000. Phone: (301) 594-5283; Fax: (301) 480-3281. E-mail:pmeltzer@nhgri.nih.gov

expression profiles associated with histological grade, DNA ploidy, and progesterone receptor (PgR) status in these tumors. Ours and similar studies may give us a better understanding of the underlying biological events in tumors that display these different clinical properties and may one day be used to augment presently used laboratory evaluation of breast cancer.

Materials and Methods

Tissues and Cells

Macroscopically fresh primary tumors from 48 node-negative breast cancer patients, tumor size 20–50 mm, were collected in the South Sweden Health Care Region (Supplemental Table 1).[5] Microscopic examination of touch preparations verified the presence of cancer cells in all samples. Nineteen of the tumors were included in a previous study (4). Steroid receptor protein (ER and PgR) determinations using enzyme immunoassay (8), as well as the flow cytometric analysis of SPF and ploidy status (9), were performed using standard methods in the routine clinical laboratory. The evaluation and interpretation of obtained data from the flow cytometer was performed according to published guidelines in cytometry (10) and instructions for the ModFitLT 3.1 software. SPF was not calculated if the histogram showed a debris distribution pattern grossly influencing the S-phase region or the fraction of nondiploid cells was <15% of all observed histogram events. Furthermore, SPF was not reported for cell populations lacking a visible G_2 peak or for which the coefficient of variation of the G_0/G_1 peak exceeded 8%. Histological grade was reevaluated by one of us (G. C.) according to Elston and Ellis (11). The grading procedure consisted of evaluation of tubule formation, nuclear pleomorphism, and mitotic count. Each of these morphological features was given a score of 1–3 points. The overall histological grade was obtained by adding these points and was categorized as follows: grade 1, 3–5 points; grade 2, 6–7 points; and grade 3, 8–9 points.

Microarray Analysis

cDNA microarray analysis was performed as described previously (4, 12) and according to standard protocols (http://research.nhgri.nih.gov/microarray/protocols.html). In short, 200 μg of BT-474 total RNA and 65–100 μg of tumor total RNA were used to produce labeled cDNA by anchored oligo(deoxythymidylate)-primed reverse transcription using SuperScript II reverse transcriptase (Invitrogen, Carlsbad, CA) in the presence of either Cy5-dUTP or Cy3-dUTP (Amersham Pharmacia, Piscataway, NJ), respectively. The arrays used were spotted with 6728 sequence-verified cDNA clones obtained from Research Genetics (Invitrogen). Fluorescence scanning and image analysis with DeArray software were performed as described previously (13, 14).

Data Analysis

For each gene, the expression intensity of the most intense channel (red or green) for each sample was averaged over all samples. All genes for which this average exceeded 300 fluorescence units (scale 0–65,535 units) were included in the analysis. In addition, we required, for all samples, that the red and green intensities both exceeded 20 fluorescence units and that the union (of the two channels) spot area exceeded 30 pixels. These requirements left us with different fractions of the original 6728 genes for the different classification problems, depending on the samples included in the analysis, which in turn was determined by the availability of measured clinical variables to be predicted (ER value and DNA ploidy: 48 samples leaving 3855 genes; PgR: 47 samples leaving 3880 genes; SPF: 45 samples leaving 3924 genes; histological grade: 35 samples leaving 4054 genes).

The data analysis was an extension of what was used by Khan *et al.* (15) and Gruvberger *et al.* (4). In brief, principal component analysis projections of the gene expression data were used as inputs to ANNs, and a classifier consisting of a committee of networks was obtained using a 3-fold cross-validation scheme. An ANN sensitivity measure was used to determine the importance of individual genes for the classification. Three extensions to this procedure were introduced: (*a*) ''cross-testing'' for better statistics in the test results; (*b*) a systematic search for the best ANN design; and (*c*) application to regression problems.

Cross-Testing. The predictive power of a committee can be tested by applying the committee to blind tests. Khan *et al.* (15) and Gruvberger *et al.* (4) used fixed blind test sets. In the present study, this was extended, for better statistical significance, to a 7-fold ''cross-testing'' procedure analogous to a cross-validation scheme (see supplemental methods). Each ANN committee was thus based on 6 of 7 available samples. With the 3-fold cross-validation procedure, each ANN model was then trained on (2/3) * (6/7) = 4/7 of the available samples. With this cross-testing, we obtained as many test results as there were samples. The cross-testing was repeated five times. Thus, the blind test result for a sample was the average result of five different committees.

ANN Architecture Selection. To obtain ANN committees with good predictive power, the ANN designs, architecture, and training parameters as described by Khan *et al.* (15) were selected to optimize the validation result [in terms of mean squared error (MSE)]. To avoid information leaks in the cross-testing scheme, every member of a predefined pool of different ANN designs was considered for each new blind test selection.

Regression Problems. Part of the analyses involved regression problems (*i.e.*, prediction of continuous values such as ER protein expression levels rather than binary classifications). For regression problems, no logistic response function was applied in the ANN output layer, and the output was directly associated with the target value. As a measure of the performance, the MSE normalized with the variance (Var) of measured (target) values was used. With this normalization, comparisons between the regression problem performances can be made. If there is no

[5]Supplementary data for this article are available at *MCT* Online (http://mct.aacrjournals.org).

useful information in the ANN inputs, MSE/Var = 1, while MSE/Var < 1 indicates a meaningful prediction. Furthermore, it is possible to evaluate the statistical significance of MSE/Var < 1 (for details, see Supplement).[5]

Gene Lists. Based on the committee of trained networks, the genes were ranked using a sensitivity measure similar to that of Khan *et al.* (15), although with a few modifications. The new sensitivity definition for a gene was based on the partial derivatives of the ANN output layer arguments, with respect to the gene expression. For each sample, these derivatives were averaged over ANN models, and the absolute value of these committee averages was then averaged over samples to get the sensitivity. Motivations for this sensitivity are given in the supplement. The analysis steps above were then redone using only the 100 genes with highest sensitivity. Note that for each choice of test set, a different gene list was used. To better evaluate the statistical significance of a high sensitivity measure of a gene, a permutation test was performed to calculate the probability α that a gene gets a larger sensitivity in a problem where target values are randomly permuted. This permutation analysis is further described in the supplemental methods.

In principle, it is possible to combine the different gene lists to one single list, but it would be computationally very costly to generate gene lists in this way in a permutation test. Instead, the most frequently generated ANN design was chosen, and a committee of 600 nets trained on different subsets of all available samples was employed, using 3-fold cross-validation.

Molecularly Motivated ER Cutoff. We investigated the possibility to define an ER protein concentration cutoff from gene expression profiles. Classification into ER positives and ER negatives, based on gene expression levels, was done for every possible partition (from having only samples with ER protein concentration = 0 as ER negatives to having only two samples with the largest available ER concentration, ≥490 fmol/mg protein, as ER positives) and the success of the classification was used as a measure of how well the partition corresponds to molecularly distinct classes. Fisher's linear discriminant (16) was used as a classifier in this analysis.

To distinguish the classification performance of different class partitions, a leave-one-out cross-validation was performed, and the area under the receiver operating characteristic (ROC) curve (17) was calculated based on the validation results. Different choices of the decision threshold correspond to different balances between the sensitivity and the specificity of the classification. All possible thresholds cause the ROC curve in the (sensitivity, 1 − specificity) plane. The area under this curve (ROC area) is a convenient measure of the classification performance with a greater area (closer to 1) signifying better performance. In Fisher's discriminant analysis, the samples are projected down to one dimension, and to compare validation results based on different projections, the scale of the one-dimensional projection result was fixed by setting the mean of the ER-negative and ER-positive classes to −1 and 1, respectively.

Results

Prediction of Continuous Values of ER Protein Expression

Gene expression profiles of 48 breast cancer specimens were determined on 6728-element cDNA microarrays. We extended our previously reported ANN methods (15) by incorporating ''cross-testing'' and applying them to regression problems to predict ER protein levels. ANN committees were constructed, using all 3855 genes that survived the filtering for quality, to predict ER values of blind tests (see Materials and Methods). From this result, the genes were ranked according to their importance for the prediction. Subsequently, the procedure was redone including only the top 100 ranked genes (see Supplemental Table 2).[5] The resulting ANN committee prediction of ER values for the blind test was compared with the protein expression values and is shown in Fig. 1. The performance of the prediction was good (MSE/Var = 0.28, $P = 1 \times 10^{-14}$; see Materials and Methods), and apart from a few tumors, the prediction of the ER values follows and correlates well with the measured protein expression values. The gene list generated from ER protein value predictions showed significant similarities among the top ranked genes compared with the gene list generated from ER status prediction of the same data (60% for the top 100 genes). The complete gene list is available in Supplemental Table 2.[5]

Prediction of the Continuous Values of ER Protein Expression When Excluding ER and Other Top Discriminators

We investigated the degree of dissimilarity between tumors displaying a higher ER protein expression and tumors with a lower ER value by estimating the number of genes that contributed to successful continuous value predictions of ER. We used a slightly modified approach compared with our previous study (4) and successively removed varying numbers of genes (3, 10, 30, 100, 300, and

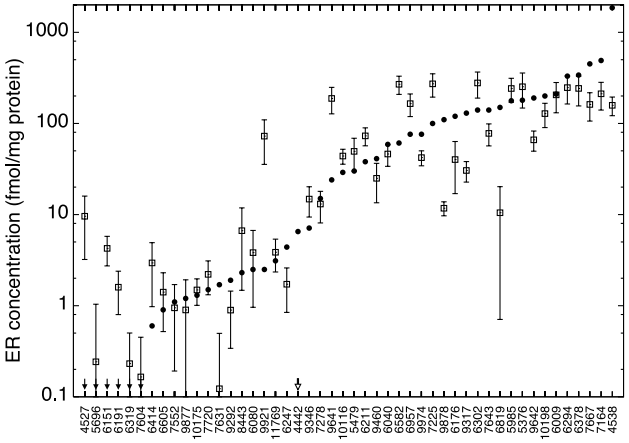

Figure 1. Results from ANN committee predictions of ER values using the top 100 ranked genes. Only the results from the blind tests are shown (*open squares*). *Solid diamonds,* measured protein expression values of the tumors; *numbers,* tumors. For this prediction, MSE/Var = 0.28 ($P = 1 \times 10^{-14}$), illustrating an excellent prediction performance.

1000) from the top of our ranked gene list and used all remaining genes or only the top 100 of the remaining genes for predictions. As expected, MSE/Var rises with the number of removed genes when only the remaining top 100 genes are used (Fig. 2), showing that genes further down on the list carry less information about ER protein expression. Still, the prediction performance is adequate and significantly better than random even when removing as many as the top 1000 genes (MSE/Var = 0.69, $P = 1.5 \times 10^{-4}$). In contrast, when using all remaining genes, almost no deterioration with number of removed genes is observed (Fig. 2). Thus, even when removing the top 1000 genes, the remaining 2855 genes carry enough information about ER signaling for a robust prediction to occur (MSE/Var = 0.38,$P = 1 \times 10^{-7}$). As can be seen in Fig. 2, while the prediction using only the top 1–100 genes yields the best performance (MSE/Var = 0.28,$P = 1 \times 10^{-14}$), it is obvious that many more genes contribute to the characteristic gene expression profile associated with ER status.

A Molecularly Defined Cutoff Value for Discriminating ER-Positive from ER-Negative Tumors

We hypothesized that ER status could be assigned using biological characteristics of tumors such as gene expression profiles. We tested this by determining the optimal distribution of tumors into the two groups, ER+ and ER−, with the best prediction performance for ER status prediction. In other words, the distribution with the most obvious separation (best prediction performance) between the two groups selects the best cutoff value for distinguishing ER+ tumors from ER− tumors. The highest prediction performance was found at an ER cutoff in the range of 6.5–15 fmol/mg protein (Fig. 3). This peak in prediction performance was seen when using the top 100 genes generated from ER value prediction as well as when using all genes that survived the filtering. Only three of the tumors included in this study had protein expression values within this range (with 6.5, 7.1, and 15 fmol/mg protein, respectively), so a more exact cutoff value could not be generated from this sample set. An ANN prediction of ER status was performed using this new cutoff threshold. When using the top 100 genes, the ANN committee results for the blind tests show that 4 of 48 tumors are misclassified. Two of these are tumors that display ER protein values around the cutoff threshold (7.1 and 15 fmol/mg protein, respectively).

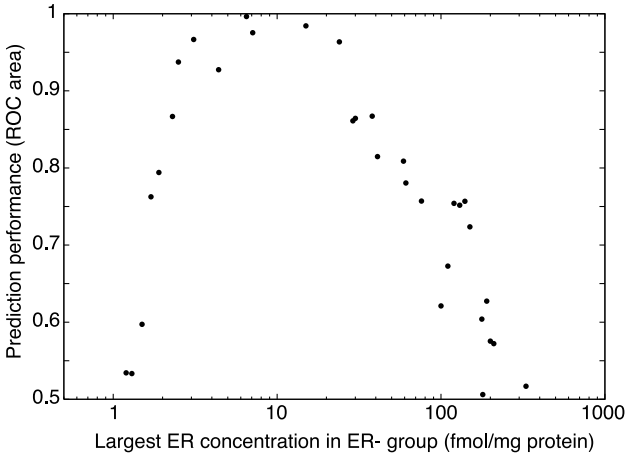

Figure 3. Determining an ER protein cutoff using the expression profiles of breast tumors. The performance, in terms of the ROC area, of a leave-one-out cross-validation scheme using Fisher's linear discriminant (see Materials and Methods) is shown for all possible partitions of the data set into ER negatives and ER positives. Best performance (largest ROC area) is seen for an ER cutoff between 6.5 and 15 fmol/mg protein. Results are shown for analyses based on the top 100 ER-associated genes (*solid circles*), but a peak was seen in the same interval for analysis using all genes that survived the filtering.

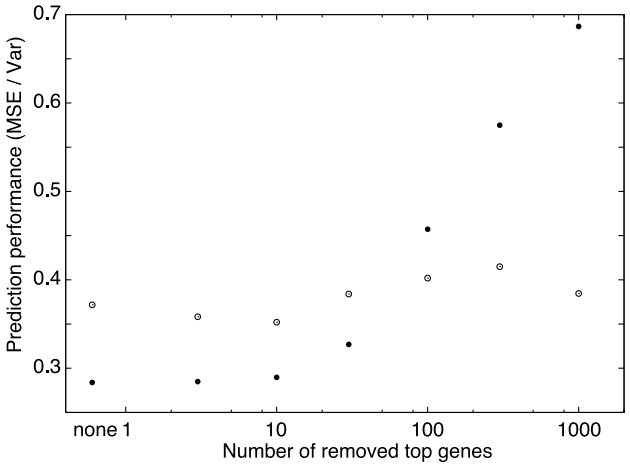

Figure 2. ER value prediction using a variable number of genes to illustrate the large number of genes influencing the prediction. Genes from the top of the ranked gene list generated from ER protein expression value prediction are removed. *Open circles,* all remaining genes are used for the prediction; *solid circles,* only top 100 among the remaining genes are used. The prediction performance is calculated as MSE/Var.

Prediction of SPF

Using the same approach as with ER values, the S-phase fraction for each tumor was also predicted from the gene expression data. The performance for the prediction of SPF for the blind tests was significant and robust (MSE/Var = 0.80, $P = 8.0 \times 10^{-4}$) when using all genes that survived the filtering. This indicates that the SPF value is encoded within the global gene expression profile of a tumor. There is a correlation between high S phase and ER negativity in the sample set studied (69% of all the ER-positive tumors have a low SPF). While there was some influence of ER on SPF prediction (when using only ER protein values as input for the ANNs, the performance was poorer: MSE/Var = 0.86, $P = 3.0 \times 10^{-3}$), the influence is not overwhelming, as the list of ranked genes for SPF value predictions deviates substantially from the list of genes for the ER value predictions. Only 20 of top 100 ranked genes for the SPF value predictions are also included in the top 100 genes from the ER value prediction (Fig. 4). Interestingly, all 20 of these genes have inverse effects on predicting ER and SPF. Remarkably, when an additional 50 genes from the top SPF predictors, which fall lower on the ER gene list, are

Rank[a] SPF	Rank[a] ER	Level[b] in high SPF	Level[c] in high ER	Gene Symbol	Gene Description	Clone ID no.
2	71			APOD	apolipoprotein D	838611
5	8			-	ESTs	155072
6	84			MYBL2	v-myb myeloblastosis viral oncogene homolog (avian)-like 2	815526
18	1			TFF3	Human Intestinal trefoil factor 3	298417
19	14			NAT1	N-acetyltransferase 1 (arylamine N-acetyltransferase)	66599
20	10			SCNN1A	sodium channel, nonvoltage-gated 1 alpha	810873
26	29			SCYA14	small inducible cytokine subfamily A (Cys-Cys), member 14	199663
27	63			CDH3	cadherin 3, type 1, P-cadherin (placental)	773301
32	7			-	ESTs	111389
40	5			CRIP1	cysteine-rich protein 1 (intestinal)	1323448
42	96			LYZ	lysozyme (renal amyloidosis)	293925
43	85			PSMB5	proteasome (prosome, macropain) subunit, beta type, 5	1460110
47	9			CEACAM6	carcinoembryonic antigen-related cell adhesion molecule 6 (non-specific cross reacting antigen)	509823
62	82			MMP9	matrix metalloproteinase 9 (gelatinase B, 92kD gelatinase, 92kD type IV collagenase)	22040
65	93			DDX11	DEAD/H (Asp-Glu-Ala-Asp/His) box polypeptide 11 (CHL1-like helicase homolog, S. cerevisiae)	741841
68	83			C1orf29	chromosome 1 open reading frame 29	754479
80	18			SLC9A3R1	solute carrier family 9 (sodium/hydrogen exchanger), isoform 3 regulatory factor 1	773286
88	21			EGR3	early growth response 3	26568
92	75			CDC20	CDC20 cell division cycle 20 homolog (S. cerevisiae)	898062
94	91			LAD1	ladinin 1	121551

Figure 4. Genes found among the top 100 ranked genes for both S-phase fraction predictions and ER value predictions using gene expression data from breast tumors. Genes are ranked based on their importance for the classification (the sensitivity value) for the different predictions, SPF and ER. *Red,* a gene is relatively overexpressed; *green,* a gene is relatively underexpressed in high SPF or high ER. The numerical position of the genes in the respective gene lists is indicated. It is interesting to note the consistent inverse relationship of the expression levels of each gene in the SPF and ER prediction gene lists. [a]The genes are ranked according to the sensitivity analysis (see Materials and Methods). [b]*Red,* higher expression in tumors with higher SPF; *green,* higher expression in tumors with low SPF. Defined as the sign of the ANN sensitivity. [c]*Red,* higher expression in tumors with higher ER; *green,* higher expression in tumors with low ER. Defined as the sign of the ANN sensitivity.

examined, 43 of 50 show an inverse relationship for predicting these two properties of breast cancer (Fig. 5). The complete ranked gene list for SPF prediction is available in Supplemental Table 3.[5]

Prediction of Other Prognostic Markers

We attempted to predict other clinically used prognostic markers such as histological grade, DNA ploidy status, and PgR protein level. Because of uneven distribution in the different categories, histological grade was analyzed as two categories, class I and class II combined *versus* class III. When predicting histological grade, 37% (13 of 35) of the tumors were misclassified (ROC area = 0.69). Because ER-associated genes might contribute to the prediction as there is a correlation between high histological grade and ER status, we predicted histological grade using the ER protein values as input for the ANNs. Indeed, the same number of tumors was misclassified (13 of 35), although with a slightly lower ROC area (0.61). The identities of the 13 misclassified tumors in the two predictions were not completely the same (8 were overlapping). We found that predicting the DNA ploidy status, diploid or nondiploid, from the gene expression profiles gave a performance with 38% (18 of 48) of the tumors misclassified (ROC area = 0.60). This indicates that tumor ploidy is not directly correlated to any specific molecular characteristics and hence has no specific stereotypic gene expression profile. Prediction of PgR protein values using all tumors indicated a good prediction performance (MSE/Var = 0.61, $P = 1.0 \times 10^{-6}$). However, there was a strong influence of ER in the prediction (ER protein expression values used as input instead of gene expression data gave MSE/Var = 0.66, $P = 5.0 \times 10^{-6}$). This is not surprising because all ER-negative tumors also were PgR negative. It would have been interesting to predict PgR values within the group of ER-positive tumors, but this was not possible due to a low number of tumors.

Discussion

The ability to predict the biological behavior of breast tumors enables selection of the optimum treatment and follow-up strategies. Although the prognostic and treatment-predictive markers presently in use in breast cancer management provide valuable information, they are not fully adequate in identifying the cancers that require more therapy or determine the most optimal therapy for the individual patient. To study the biology behind some of the prognostic markers presently in use, the expression of 6728 genes was investigated in primary tumor tissues from 48 breast cancer patients. The tumors came from a well-characterized group of node-negative breast cancers.

The ER status of a tumor is determined from its protein value and has long been used as a means to identify the group of patients that will benefit from endocrine therapy. However, the ER status based on protein expression does not give a direct verification of the functional activity in the ER signaling pathways. In previous studies of global gene expression of breast tumors, it has become evident that the ER status of tumors is associated with distinct gene expression profiles involving a large number of genes (4–6). However, these studies have only focused on the binary ER status and did not examine the relationship of gene expression profiles to the continuous range of ER

Rank[a] SPF	Rank[a] ER	Level[b] in high SPF	Level[c] in high ER	Gene Symbol	Gene Description	Clone ID no.
1	790			HMGCS2	3-hydroxy-3-methylglutaryl-Coenzyme A synthase 2 (mitochondrial)	757222
3	2555			IGHG3	immunoglobulin heavy constant gamma 3 (G3m marker)	814124
4	1357			FOS	v-fos FBJ murine osteosarcoma viral oncogene homolog	811015
7	1738			AZGP1	alpha-2-glycoprotein 1, zinc	1456160
8	1621			FCGR3B	Fc fragment of IgG, low affinity IIIb, receptor for (CD16)	51447
9	2965			SLC6A8	solute carrier family 6 (neurotransmitter transporter, creatine), member 8	725877
10	1957			EPHB3	EphB3	813520
11	1053			MCAM	melanoma cell adhesion molecule	897531
12	1241			UBE2C	ubiquitin-conjugating enzyme E2C	769921
13	491			PLIN	perilipin	108330
14	1015			GCHFR	GTP cyclohydrolase I feedback regulatory protein	525799
15	141			INHBA	inhibin, beta A (activin A, activin AB alpha polypeptide)	269815
16	236			-	Homo sapiens, clone IMAGE:3881549, mRNA	143169
17	620			PTN	pleiotrophin (heparin binding growth factor 8, neurite growth-promoting factor 1)	361974
21	2212			SCYA28	CC chemokine CCL28	136919
22	349			ALCAM	activated leucocyte cell adhesion molecule	26617
23	2242			IGF1	insulin-like growth factor 1 (somatomedin C)	813179
24	140			TFAP2B	transcription factor AP-2 beta (activating enhancer binding protein 2 beta)	363144
25	3205			LTF	lactotransferrin	460487
28	803			DCT	dopachrome tautomerase (dopachrome delta-isomerase, tyrosine-related protein 2)	753104
29	2240			ADAM8	a disintegrin and metalloproteinase domain 8	704254
30	779			TCN1	transcobalamin I (vitamin B12 binding protein, R binder family)	592243
31	525			TMSNB	thymosin, beta, identified in neuroblastoma cells	306771
33	492			IFI30	interferon, gamma-inducible protein 30	856447
34	471			TEK	TEK tyrosine kinase, endothelial (venous malformations, multiple cutaneous and mucosal)	151501
35	485			FABP4	fatty acid binding protein 4, adipocyte	307660
36	972			STMN1	stathmin 1/oncoprotein 18	1476065
37	389			RCN2	reticulocalbin 2, EF-hand calcium binding domain	898253
38	2437			SPS	selenium donor protein	840702
39	480			RUNX3	runt-related transcription factor 3	291478
41	194			JPO1	Cell division cycle associated 7, c-Myc target JPO1	244058
44	2434			FYB	FYN binding protein (FYB-120/130)	293325
45	2445			COL8A1	collagen, type VIII, alpha 1	1472775
46	338			-	ESTs, Weakly similar to T33068 hypothetical protein C35E7.9 - Caenorhabditis elegans	131316
48	2709			ATF3	activating transcription factor 3	51448
49	360			SLU7	step II splicing factor SLU7	80948
50	3571			RBBP1	retinoblastoma binding protein 1	502832
51	800			PLIN	perilipin	108330
52	243			MIG	monokine induced by gamma interferon	503617
53	2898			MMP11	matrix metalloproteinase 11 (stromelysin 3)	487296
54	835			SLC16A1	solute carrier family 16 (monocarboxylic acid transporters), member 1	486175
55	498			SEPP1	selenoprotein P, plasma, 1	530814
56	1144			NDRG1	N-myc downstream regulated gene 1	842863
57	3338			BGN	biglycan	244147
58	118			DUSP4	dual specificity phosphatase 4	756596
59	529			RSU1	Ras suppressor protein 1	687397
60	502			NR4A2	nuclear receptor subfamily 4, group A, member 2	898221
61	1980			COL11A1	collagen, type XI, alpha 1	134783
63	3626			FLJ14146	hypothetical protein FLJ14146	131887
64	408			BUB1	BUB1 budding uninhibited by benzimidazoles 1 homolog (yeast)	781047

Figure 5. Genes found among the top 100 ranked genes for the S-phase fraction predictions but not among the top 100 for ER value predictions using gene expression data from breast tumors. Genes are ranked based on their importance for the classification (the sensitivity value) for the different predictions, SPF and ER. *Red,* a gene is up-regulated for high SPF or high ER; *green,* a gene is down-regulated. The positions of the genes in the respective gene lists are indicated. [a]The genes are ranked according to the sensitivity analysis (see Materials and Methods). [b]*Red,* higher expression in tumors with higher SPF; *green,* higher expression in tumors with low SPF. Defined as the sign of the ANN sensitivity. [c]*Red,* higher expression in tumors with higher ER; *green,* higher expression in tumors with low ER. Defined as the sign of the ANN sensitivity.

protein values. In this study, we successfully calculated the ER protein expression values from gene expression profiles, showing that gene expression data from tumors are sufficiently robust and informative not only to determine ER status but also to indicate the actual level of ER protein expression. Moreover, the strength of the ER profile is evidenced when even after removing the most important 1000 genes of the ER profile, we were still able to predict the ER protein values with good performance (MSE/Var = 0.69, $P = 1.5 \times 10^{-4}$; Fig. 2). The genes associated with ER protein expression value predictions are to a large extent overlapping with the genes associated with ER status prediction in this and other studies (2,4–6). Conventionally, the threshold value used to assign ER status (positive or

negative) has been determined empirically from response to endocrine treatment, and the cutoffs used differ between laboratories and clinics (18). Using the ER-associated gene expression profiles, we have determined a protein level cutoff for ER status. In this patient cohort, an appropriate cutoff for ER status based on the top 100 ER-associated genes (from the continuous value predictions) is in the range of 6.5–15 fmol/mg protein. Only few tumors were within this range of protein expression values, which therefore was difficult to narrow. Still, this range of values is somewhat lower than the cutoff that was used at the hospitals of origin of the tumors at the time of diagnosis for these patients (25 fmol/mg protein). Because the number of samples in this study, especially in the critical range, is limited, this cutoff value may not be applicable to other patient cohorts. However, this approach appears sufficiently promising to warrant studies with larger numbers of tumors. Determining an ER status cutoff threshold based on the expression of a panel of genes associated with ER in breast tumors could possibly be a more accurate way of assigning their ER status than using merely the ER protein level.

The proliferative activity of a tumor can be estimated by flow cytometric analysis whereby information on DNA ploidy status and SPF is generated. We found that the performance for the prediction of SPF values based on gene expression profiles is good. It should be mentioned that because of the correlation between ER status and SPF in the patient cohort, the strong signal from genes associated with ER status contributes to some degree to the prediction of SPF. However, a low overlap of the top ranked genes between the S-phase- and ER-associated gene expression profiles (20%) indicates that although ER-associated genes do assist in the prediction of SPF, most genes important to SPF prediction are indeed associated more specifically with the S-phase profile. Interestingly, all of the 20 genes comprising the intersection of the top 100 SPF and top 100 ER value gene lists show an inverse relation in their expression in that the genes that are highly expressed in tumors with high S-phase fraction have a low expression in tumors with high ER values (Fig. 4). Additionally, 67 of the remaining 80 genes on the top 100 S-phase list that are found far down the ER value gene list also show an inverse correlation in expression level (Fig. 5 and Supplemental Tables 2 and 3).[5] This striking inverse relationship shows that at the molecular level, the gene expression of many individual genes important to a high proliferation phenotype relates directly to a low expression of ER protein. Not surprisingly, several of the 80 genes that are strongly associated with SPF but not with ER have functions associated with cell proliferation. For example, the ubiquitin-conjugating enzyme E2C (rank 12) is highly expressed in tumors with a high S-phase fraction and involved in the ubiquitin-dependent proteolysis of both cyclin A and cyclin B (19). The cell growth-inhibiting transcription factors AP-2β (rank 24; Ref. 20) and activating transcription factor 3 (rank 48; Ref. 21) both have a low expression in tumors with high S phase as do the inhibin βA subunit of the inhibin complex (rank 15; Ref. 22) and insulin-like growth factor-1 (rank 23; Ref. 23), both are involved in modulation of cell growth and differentiation. Another group of genes associated with high S-phase fraction are genes that have previously been associated with tumor invasion, tumorigenesis, and transformation including ADAM8 (a disintegrin and metalloproteinase domain 8; rank 29), a transmembrane protein identified to have metalloproteinase activity (24), and melanoma cell adhesion molecule (rank 11), which has been implicated to play an important role in initiation and malignant progression in melanoma and prostate cancer (25). The fact that the gene transcobalamin I was more highly expressed in tumors with a low S phase suggests that a higher S-phase fraction in a tumor also seems to be correlated with a lower degree of cellular differentiation. Transcobalamin I, a member of the vitamin B12 binding protein family also called R binders, has been demonstrated by immunohistochemistry to be expressed more often in the well-differentiated tumors in invasive ductal carcinomas of the breast (26).

The histological grade of a tumor is determined by microscopic evaluation of breast tumor paraffin sections. From our results, predicting histological grade from gene expression profiles seems to be possible. Although there is an influence by ER status, owing to a correlation between low histological grade and ER positivity, the ER protein expression values alone predict the histological grades less accurately than the gene expression profiles. Although their study did not address the prediction of histological grade, very recently, gene expression profiles have been observed which distinguish high- and low-grade tumors (27).

The DNA ploidy status of a tumor can reveal whether some cells in the tumor have an abnormal amount of DNA in the nucleus. The prediction of ploidy status (diploid *versus* nondiploid) was not as good as for the other clinical parameters studied, indicating that the DNA ploidy status of a tumor is not strongly correlated to any specific gene expression profile. It is not surprising that it was difficult to find a unifying gene expression profile for all nondiploid tumors because their chromosomal gains and losses do not necessarily follow the same pattern; therefore, the effects of aneuploidy on gene expression are diverse. Possibly, better results could be obtained by grouping tumors according to comparative genomic hybridization profiles, which are determined by specific patterns of cytogenetic change. Indeed, several studies have reported correlations between comparative genomic hybridization profiles and gene expression data (28–30).

Our study sheds light on the molecular background behind the already established markers ER status and SPF. Using computer models, we were able to predict the continuous values of these clinically relevant markers, demonstrating that the biological basis of these markers is encoded and detectable within global gene expression patterns, even from within heterogeneous tumor samples. The method of predicting a tumor characteristic on a continuous scale may be a better approach than predicting binary classes in other microarray studies (*e.g.*, prediction of time to disease recurrence instead of recurrence by a

fixed end point). Additional studies and a reliable approach to generate expression data in a clinical setting are necessary before gene expression profiling can be used as a practical clinical tool. However, our study and others strongly suggest the approaching potential of gene expression profiling to aid treatment decision-making for the individual patient by refining prognostic categories and elucidating the molecular properties that affect outcome.

Acknowledgments

We thank Thomas Breslin for technical assistance and the participating departments of the South Sweden Breast Cancer Group for providing us with breast cancer samples.

References

1. Martin KJ, Kritzman BM, Price LM, et al. Linking gene expression patterns to therapeutic groups in breast cancer. Cancer Res, 2000;60: 2232 – 8.

2. Perou CM, Sorlie T, Eisen MB, et al. Molecular portraits of human breast tumors. Nature, 2000;406:747 – 52.

3. Sorlie T, Perou CM, Tibshirani R, et al. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proc Natl Acad Sci USA, 2001;98:10869 – 74.

4. Gruvberger S, Ringner M, Chen Y, et al. Estrogen receptor status in breast cancer is associated with remarkably distinct gene expression patterns. Cancer Res, 2001;61:5979 – 84.

5. West M, Blanchette C, Dressman H, et al. Predicting the clinical status of human breast cancer by using gene expression profiles. Proc Natl Acad Sci USA, 2001;98:11462 – 7.

6. van't Veer LJ, Dai H, van de Vijver MJ, et al. Gene expression profiling predicts clinical outcome of breast cancer. Nature, 2002;415:530 – 6.

7. van de Vijver MJ, He YD, van't Veer LJ, et al. A gene-expression signature as a predictor of survival in breast cancer. N Engl J Med, 2002;347:1999 – 2009.

8. Ferno M, Stal O, Baldetorp B, et al. Results of two or five years of adjuvant tamoxifen correlated to steroid receptor and S-phase levels. South Sweden Breast Cancer Group, and South-East Sweden Breast Cancer Group. Breast Cancer Res Treat, 2000;59:69 – 76.

9. Baldetorp B, Dalberg M, Holst U, Lindgren G. Statistical evaluation of cell kinetic data from DNA flow cytometry (FCM) by the EM algorithm. Cytometry, 1989;10:695 – 705.

10. Baldetorp B, Bendahl PO, Ferno M, et al. Reproducibility in DNA flow cytometric analysis of breast cancer: comparison of 12 laboratories' results for 67 sample homogenates. Cytometry, 1995;22:115 – 27.

11. Elston CW, Ellis IO. Pathological prognostic factors in breast cancer. I. The value of histological grade in breast cancer: experience from a large study with long-term follow-up. Histopathology, 1991;19:403 – 10.

12. Khan J, Simon R, Bittner M, et al. Gene expression profiling of alveolar rhabdomyosarcoma with cDNA microarrays. Cancer Res, 1998;58: 5009 – 13.

13. Khan J, Bittner ML, Chen Y, Meltzer PS, Trent JM. DNA microarray technology: the anticipated impact on the study of human disease. Biochim Biophys Acta, 1999;1423:M17 – 28.

14. Chen Y, Dougherty ER, Bittner ML. Ratio-based decisions and the quantitative analysis of cDNA microarray images. Biomed Optics, 1997; 364 – 74.

15. Khan J, Wei JS, Ringner M, et al. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nat Med, 2001;7:673 – 9.

16. Fisher RA. The use of multiple measurements in taxonomic problems. Ann Eugen, 1936;7:179 – 88.

17. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology, 1982;143: 29 – 36.

18. Harvey JM, Clark GM, Osborne CK, Allred DC. Estrogen receptor status by immunohistochemistry is superior to the ligand-binding assay for predicting response to adjuvant endocrine therapy in breast cancer. J Clin Oncol, 1999;17:1474 – 81.

19. Townsley FM, Aristarkhov A, Beck S, Hershko A, Ruderman JV. Dominant-negative cyclin-selective ubiquitin carrier protein E2-C/UbcH10 blocks cells in metaphase. Proc Natl Acad Sci USA, 1997;94:2362 – 7.

20. Zeng YX, Somasundaram K, el-Deiry WS. AP2 inhibits cancer cell growth and activates p21WAF1/CIP1 expression. Nat Genet, 1997;15:78 – 82.

21. Chen BP, Liang G, Whelan J, Hai T. ATF3 and ATF3 δ zip. Transcriptional repression *versus* activation by alternatively spliced isoforms. J Biol Chem, 1994;269:15819 – 26.

22. Jiang X, Russo IH, Russo J. Human chorionic gonadotropin and inhibin induce histone acetylation in human breast epithelial cells. Int J Oncol, 2002;20:77 – 9.

23. Furstenberger G, Senn HJ. Insulin-like growth factors and cancer. Lancet Oncol, 2002;3:298 – 302.

24. Amour A, Knight CG, English WR, et al. The enzymatic activity of ADAM8 and ADAM9 is not regulated by TIMPs. FEBS Lett, 2002;524: 154 – 8.

25. Wu GJ, Wu MW, Wang SW, et al. Isolation and characterization of the major form of human MUC18 cDNA gene and correlation of MUC18 over-expression in prostate cancer cell lines and tissues with malignant progression. Gene, 2001;279:17 – 31.

26. Ogawa K, Kudo H, Kim YC, Nakashima Y, Ohshio G, Yamabe H. Expression of vitamin B12 R-binder in breast tumors. An immunohistochemical study. Arch Pathol Lab Med, 1988;112:1117 – 20.

27. Ma X-J, Salunga R, Tuggle JT, et al. Gene expression profiles of human breast cancer progression. Proc Natl Acad Sci, 2003;100:5974 – 9.

28. Bayani J, Brenton JD, Macgregor PF, et al. Parallel analysis of sporadic primary ovarian carcinomas by spectral karyotyping, comparative genomic hybridization, and expression microarrays. Cancer Res, 2002;62: 3466 – 76.

29. Hyman E, Kauraniemi P, Hautaniemi S, et al. Impact of DNA amplification on gene expression patterns in breast cancer. Cancer Res, 2002;62:6240 – 5.

30. Pollack JR, Sorlie T, Perou CM, et al. Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors. Proc Natl Acad Sci USA, 2002;99: 12963 – 8.