SUBTYPE SPECIFIC BREAST CANCER EVENT PREDICTION

Herman Sontrop[1], *Wim Verhaegh*[1], *René van den Ham*[1], *Marcel Reinders*[2] *and Perry Moerland*[3]

[1]Philips Research, High Tech Campus 12a, 5656 AE Eindhoven, The Netherlands
[2]Delft University of Technology, Mekelweg 4, 2628 CD Delft, The Netherlands
[3]Academic Medical Center, Meibergdreef 9, 1100 AZ Amsterdam, The Netherlands
{herman.sontrop, wim.verhaegh, rene.van.den.ham}@philips.com,
m.j.t.reinders@tudelft.nl, p.d.moerland@amc.uva.nl

ABSTRACT

We investigate the potential to enhance breast cancer event predictors by exploiting subtype information. We do this with a two-stage approach that first determines a sample's subtype using a recent module-driven approach, and secondly constructs a subtype-specific predictor to predict a metastasis event within five years. Our methodology is validated on a large compendium of microarray breast cancer datasets, including 43 replicate array pairs for assessing subtyping stability. Note that stratifying by subtype strongly reduces the training set sizes available to construct the individual predictors, which may decrease performance. Besides sample size, other factors like unequal class distributions and differences in the number of samples per subtype, easily obscure a fair comparison between subtype-specific predictors constructed on different subtypes, but also between subtype specific and subtype a-specific predictors. Therefore, we constructed a completely balanced experimental design, in which none of the above factors play a role and show that subtype-specific event predictors clearly outperform predictors that do not take subtype information into account.

1. INTRODUCTION

Breast cancer *event prediction* is an important yet challenging classification problem in which we attempt to predict whether a certain type of event will happen within a given time frame or not, in our case whether a breast tumor will metastasize or not, based on gene expression data obtained from microarrays. Although initial reports on microarray based breast cancer event prediction were very optimistic, papers using improved evaluation protocols showed that only modest performances seem attainable, with a *balanced accuracy rate* (bar; the average of sensitivity and specificity) in the range of 62 to 65% [1]. Furthermore, there is increasing evidence that breast cancer is not a single disease at the molecular level, but that breast cancers constitute a diverse and heterogeneous group of diseases. Various breast cancer subtyping schemes have been described in the literature, mostly based on the *intrinsic gene list* approach from the landmark publication by Perou et al. [2], which introduced a breast cancer subtype taxonomy that classifies breast cancer as either luminal A (lumA), luminal B (lumB), basal, Her2 or normal like, based on hierarchical clustering. Although by now it is well known that over large sample groups these subtypes are associated with a difference in survival time, only few studies couple subtyping directly to breast cancer event prediction. This is not completely surprising, as in recent years several studies indicated stability issues with the intrinsic gene list approach [3, 4], while also casting doubts on the existence of the normal like tumors as a genuine breast cancer subtype [5]. Our subtyping, however, is based on a recently introduced biology inspired module-driven approach [6], yielding the four clusters lumA, lumB, basal and Her2. This method was shown to be much more robust than the intrinsic gene list approach, mainly based on observed prediction strength [7]. We confirm the stability of this subtyping scheme using a set of technical replicate pairs, which clearly indicate the stability of the subtyping.

Besides the problem of subtype instability, an obvious drawback of developing an event predictor per subtype is that stratifying the samples by subtype strongly reduces the training set available to construct each such predictor, which may decrease performance. This can partly be overcome by pooling data from multiple studies, although when pooling data from different array platforms, proper cross-platform normalization is challenging. Fortunately, over time more breast cancer studies have appeared using the same type of microarray, allowing us to construct a large compendium more easily, and thereby alleviating the sample size problem. Besides sample size, several other factors like unequal class distributions and differences between the number of samples per subtype easily obscure a fair comparison between subtype specific predictors constructed on different subtypes, but also between subtype specific predictors and predictors that do not take subtype information into account. Therefore, in order to facilitate a fair comparison we constructed a completely balanced experimental design, in which none of the above factors play a role, showing that subtype specific event-predictors clearly outperform subtype a-specific predictors.

2. DATA AND METHODS

2.1. Compendium construction

We constructed a compendium out of ten individual data sets, all measured using Affymetrix HG-u133A

GeneChips. We list them by the first author of the corresponding study and the accession number under which the expression data and further references can be found: Wang GSE2034, Yu GSE5327, Desmedt GSE7390, Schmidt GSE11121, Minn GSE2603, Loi GSE6532, Sotiriou GSE2990, Miller GSE3494, Pawitan GSE1456 and Chin E-TABM-158. Raw expression data for all studies is available at GEO, with the exception of the dataset by Chin, which is available via ArrayExpress. All processing was performed using R and Bioconductor. After removing duplicate entries (330) and removing outlier arrays (15), detected using *arrayQualityMetrics* [8], 1539 unique hybridizations remained, involving 1496 unique samples and 43 technical replicates. Raw expression data was used to generate MAS5.0 expression estimates, using *affy*, scaled to a target intensity of 600. Before pooling all expression data, for each study and each gene separately, the expression estimates were z-transformed, as suggested in [9].

2.2. Subtyping

Subtyping was performed using the module-driven approach recently introduced [6, 7], which first transforms the expression data into three module scores, related to key biological processes strongly associated with breast cancer. The three modules constitute an ER-related module, using 468 genes, a Her2-related module of 27 genes and a proliferation-related module, referred to as AURKA, involving 228 genes. After the expression data is transformed to the module scores, a Gaussian mixture model is fitted in order to determine class membership, in which the ER and Her2 module scores are used to infer the subtypes luminal, Her2 and basal, while the AURKA module is used to further divide the luminal group into a lumA and a lumB compartment. An implementation of the methodology is available in the package *genefu*. Discarding the $43 \times 2 = 86$ replicate arrays, we used the remaining $1539 - 86 = 1453$ samples to fit the subtyping model and applied it to all data, including the replicates. Note that subtyping does not involve class label information, hence this step does not introduce an information leak to predictor construction.

2.3. Event prediction

Class labels are solely based on distant metastasis free survival time, with *poor prognosis* cases (PP) having an event i.e. distant metastasis within five years, while the *good prognosis* cases (GP) did not have an event during follow-up, with a follow-up time of at least five years. These stringent criteria led to the identification of 229 PP samples and 663 GP samples, yielding a total of 892 unique samples, which are divided over the four subtypes as indicated in Table 1.

2.4. Balanced experimental design

In this work the key difference in the construction of predictors is what type of training data is offered to it, which is either subtype-specific (e.g. only samples of subtype basal) or a-specific. We will refer to these predictors as *typed predictors* and *untyped predictors*, respectively. We evaluate the assignments made by a predictor by inspecting its sensitivity (sen), specificity (spc), balanced accuracy rate (bar), overall accuracy (acc) and its area under ROC curve (auc). Some of these indicators can be very sensitive to differences in class distribution or number of samples per subtype, which can easily obscure comparisons. From Table 1 we indeed see large differences between class distributions (ratio) and number of samples per subtype (total). Hence to allow for a fair comparison, we use a balanced sub-sampling scheme that uses the same number of samples and has the same good over poor ratio for each subtype. In our compendium the maximum group size that still allows this is to sample 74 GP samples and 36 PP samples from each subtype. Besides the Her2 group, for other groups this can be done in multiple ways, therefore this strategy was repeated 100 times.

Table 1. The distribution of class labels and subtypes of the 892 samples with proper class labels. The bottom row indicates the good over poor ratio.

	lumA	lumB	basal	Her2	total
good	273	216	100	74	663
poor	42	94	57	36	229
total	315	310	157	110	892
ratio	6.5	2.8	1.8	2.1	2.9

More formally, let D represent the complete collection of 892 samples and let D_s denote the set of all samples in the compendium of subtype s. Furthermore, let $D_{s,r}$ represent the 74 GP and 36 PP sample set randomly drawn in repeat r for subtype s. For each repeat $r = 1, \ldots, 100$, and for each sample $i \in D_r = \bigcup_s D_{s,r}$, four separate predictors are being constructed. The four predictors only differ by the training data that was offered: $T^{\text{tp}}_{i,r}$, $T^{\text{un}}_{i,r}$, $T^{\text{tp+}}_{i,r}$ or $T^{\text{un+}}_{i,r}$. These sets are constructed as follows, assuming sample i is of subtype s.

Training set for the typed predictor. Typed predictors are built using only samples of the corresponding subtype: $T^{\text{tp}}_{i,r} = D_{s,r} - \{i\}$.

Training set for the untyped predictor. Untyped predictors are built using a mix of subtypes: $T^{\text{un}}_{i,r}$ is constructed by randomly drawing samples from $D_r - \{i\}$ with the same number of GP and PP cases as in $T^{\text{tp}}_{i,r}$.

Augmented training sets. For three of the four subtypes the training sets can be made larger, so to evaluate the influence of sample size we also constructed maximum sized training sets. Hence, we augment the training set of the typed predictor to $T^{\text{tp+}}_{i,r} = D_s - \{i\}$, and the training set of the untyped predictor is augmented to $T^{\text{un+}}_{i,r} = D - \{i\}$

2.5. Predictor construction

For each of the above training sets we first performed a filtering step to rule out spurious features, in which we use the present/absent calls from the MAS5.0 procedure and only selected genes for which in at least one of the GP and

PP groups the number of present calls was at least 50%. We then used moderated-t statistics, as implemented in *limma* to rank the genes and selected the top 200 features. In this work predictors are based on the nearest centroid rule (NC), as suggested in [1], which despite its simplicity is often among best in class. Note that the NC is an automatic learner that offers no direct possibility to manipulate sensitivity or specificity via parameter tuning.

2.6. Obtaining performance estimates

After all samples $i \in D_r$ have been classified by all predictors, for each of the four predictors and for each subtype s we construct a subtype specific performance estimate by considering all assignments corresponding to samples $i \in D_{s,r}$, while we also construct an overall performance by considering all assignments corresponding to samples $i \in D_r$. Results are averaged over all 100 repeats. ROC curves are generated by using the difference between the distance of a sample to each of the centroids as a continuous criterion, on which a variable threshold is set.

In essence our strategy is a leave-one-out cross validation for the typed predictors, after which we deliberately manipulate the corresponding training sets in order to investigate the influence of subtype distribution. Note that leave-one-out cross validation is a natural candidate here, as the stratification by subtype puts severe pressure on already small sample sets.

2.7. Baseline predictor

Finally, we construct a baseline predictor that classifies each sample $i \in D$ by using $D - \{i\}$ as a training set, similarly to the augmented untyped predictor, but evaluated on the subtype sets D_s and the entire compendium D rather than the restricted, balanced sets $D_{s,r}$ and D_r.

3. RESULTS

3.1. Subtyping concordance

Figure 1 presents a 2D scatterplot of the module scores corresponding to the subtype assignments for all 1539 hybridizations. It shows that 42 out of the 43 (98%) technical replicate pairs are assigned the same subtype, when restricted to luminal vs. basal vs. Her2. If also the distinction between lumA and lumB is made, 35 out of the 43 (81%) replicate pairs are assigned concordant subtypes.

3.2. Performance of typed vs. untyped predictors

The results for the typed and untyped predictors are shown in Table 2. The table clearly shows an improvement over almost all performance indicators when comparing typed to untyped predictors, although for some groups the untyped predictors still have a better sensitivity. Especially within the Her2 group the performance difference is large.

3.3. Performance using augmented training sets

Table 3 shows how much performance can be regained compared to Table 2 if all available training samples are used. For the typed predictors again almost all performance indicators improve compared to Table 2, although the lumA predictors seem to trade some sensitivity for specificity. For the untyped predictors the story is quite different. Note that the augmented training sets in this setting are substantially larger for the untyped predictors than for the typed predictors and that the former are dominated by luminal samples. A striking difference can be observed between the sensitivity and specificity of the lumA subtype compared to the other subtypes. Apparently, the augmented untyped predictors are biased to predict a good

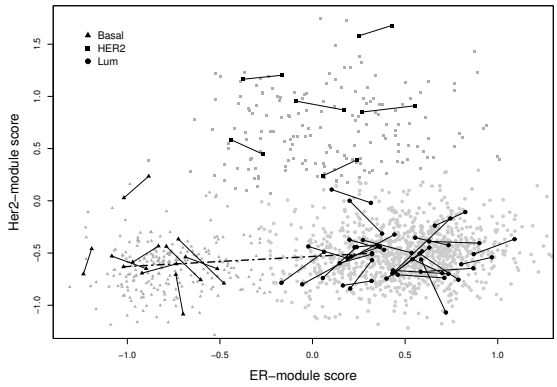

Figure 1. Module scores corresponding to the subtype assignments for the 1539 hybridizations, including the 43 technical replicate pairs (connected by line segments). The lumA and lumB groups have been collapsed, and the AURKA dimension that separates them is not shown.

Table 2. Typed (tp) vs. untyped (un) predictor performance.

		lumA	lumB	basal	Her2	overall
tp	sen	39.0	73.4	43.6	77.8	57.1
	spc	76.0	50.8	68.9	67.6	65.3
	bar	57.5	62.1	56.2	72.7	61.2
	acc	63.9	58.2	60.6	70.9	62.6
	auc	62.3	66.8	60.6	74.3	65.3
un	sen	55.1	67.9	49.4	67.1	58.7
	spc	53.1	47.2	60.4	57.7	54.2
	bar	54.1	57.6	54.9	62.4	56.4
	acc	53.8	54.0	56.8	60.8	55.7
	auc	56.4	62.3	59.0	66.2	59.4

Table 3. Typed (tp+) vs. untyped (un+) predictor performance with augmented training sets.

		lumA	lumB	basal	Her2	overall
tp+	sen	29.3	73.6	47.5	77.8	57.4
	spc	82.9	54.9	68.2	67.6	66.7
	bar	56.1	64.2	57.8	72.7	62.1
	acc	65.4	61.0	61.4	70.9	63.6
	auc	62.8	71.9	62.5	74.3	66.4
un+	sen	4.0	85.8	82.8	86.1	65.1
	spc	97.7	38.2	12.9	33.7	45.8
	bar	51.2	62.0	47.9	59.9	55.4
	acc	67.6	53.9	35.8	50.9	52.1
	auc	67.3	72.8	47.9	63.7	58.1

Table 4. Baseline predictor performance over the entire compendium.

	lumA	lumB	basal	Her2	overall
sen	4.8	86.6	82.8	86.5	70.9
spc	97.5	37.7	12.2	33.3	57.8
bar	51.2	62.1	47.5	59.9	64.4
acc	85.6	52.5	37.0	50.4	61.1
auc	67.3	73.3	46.3	64.1	69.2

prognosis for lumA samples, giving a very high specificity but very poor sensitivity for that subtype, and to predict a poor prognosis for the other subtypes, giving a high sensitivity but a rather low specificity for them. This makes sense, as from Table 1 we can observe that the number of luminal samples is large, with the vast majority of lumA samples having a good prognosis, whereas the other subtypes have relatively more poor prognosis cases. As a result, these predictors mainly separate lumA samples from the other subtypes, while within each subtype the separation between GP and PP samples is fairly weak. A closer inspection of the selected features revealed that the untyped predictors pick up a relatively large number of proliferation genes in the signature that were also in the AURKA module used to separate lumA from lumB (results not shown).

In addition, comparing Tables 2 and 3, we see that offering more samples can sometimes even decrease performance, viz. for the untyped predictors we see that versions using the augmented training sets in fact have a lower spc, bar, acc and auc for the basal and Her2 groups, which is an additional indication that these subtypes require very specific training samples.

3.4. A dissection of the baseline performance

Table 4 shows the performance of the baseline predictors. As expected, for each subtype the performance is very close to the performance of the augmented untyped predictors from Table 3. However, the column *overall* from Table 4 indicates a large improvement compared to that in Table 3. This is due to a difference in the number of samples per subtype, as the the 892 sample collection is unbalanced and dominated by the luminal subtypes, which constitute approximately 70% of the total.

We also note a very peculiar effect of the bar indicator. As we can see the bar over the complete compendium is 64.4, however, for *every* subtype the bar is less, even though the subtypes form a partition of the compendium. This is an unwanted and maybe unexpected effect of combining sets with different ratios of good over poor prognosis cases. It can easily be shown that in a bar score the good over poor ratio determines the weight of a misclassification of a poor prognosis case compared to a good prognosis case. Since the subtypes have different ratios of good over poor samples, the same misclassification is weighted very differently within each particular subtype as compared to the entire sample set.

4. CONCLUSION

In this paper we studied the role of subtyping in breast cancer event prediction. Our subtyping is based on a recently introduced module-driven subtyping approach for which an additional stability analysis was performed using a set of technical replicates.

Using a balanced sub-sampling scheme we have shown that typed predictors offer a better separation between GP and PP cases within subtypes than untyped predictors. Our results indicate a strong relation between the distribution of subtypes present in training data and the event predictions made, and especially Her2 samples seem to benefit from using a typed predictor.

Furthermore, we provided additional insight in the performance of untyped NC-based breast cancer event predictors, which mainly seem to separate the lumA subtype from the remaining subtypes, while also showing that having more samples does not necessarily improve performance over all subtypes. Finally, we have shown an unwanted effect of using the bar indicator, which shows that bar scores should not be compared when class distributions are different.

5. REFERENCES

[1] L.F.A. Wessels, M.J.T. Reinders, A.A.M. Hart, et al., "A protocol for building and evaluating predictors of disease state based on microarray data," *Bioinformatics*, vol. 21, no. 19, pp. 3755–3762, 2005.

[2] C.M. Perou, T. Sørlie, M.B. Eisen, et al., "Molecular portraits of human breast tumours," *Nature*, vol. 406, no. 6797, pp. 747–752, 2000.

[3] L. Lusa, L.M. McShane, J.F. Reid, et al., "Challenges in projecting clustering results across gene expression profiling datasets," *Journal of the National Cancer Institute*, vol. 99, no. 22, 2007.

[4] B. Weigelt, A. Mackay, R. A'hern, et al., "Breast cancer molecular profiling with single sample predictors: a retrospective analysis," *Lancet oncology*, vol. 11, no. 4, pp. 339–349, 2010.

[5] J.S. Parker, M. Mullins, M.C.U. Cheang, et al., "Supervised risk predictor of breast cancer based on intrinsic subtypes," *Journal of Clinical Oncology*, vol. 27, no. 8, pp. 1160, 2009.

[6] C. Desmedt, B. Haibe-Kains, P. Wirapati, et al., "Biological processes associated with breast cancer clinical outcome depend on the molecular subtypes," *Clinical Cancer Research*, vol. 14, no. 16, pp. 5158, 2008.

[7] B. Haibe-Kains, *Identification and assessment of gene signatures in human breast cancer*, Ph.D. thesis, University Libre de Bruxelles, Bioinformatics Department, 2009.

[8] A. Kauffmann, R. Gentleman, and W. Huber, "arrayQualityMetrics–a bioconductor package for quality assessment of microarray data," *Bioinformatics*, vol. 25, no. 3, pp. 415, 2009.

[9] H. Yasrebi, P. Sperisen, V. Praz, and P. Bucher, "Can survival prediction be improved by merging gene expression data sets," *PloS ONE*, vol. 4, no. 10, pp. e7431, 2009.