Do microarrays improve breast cancer prognosis?
A long story short

Benjamin Haibe-Kains$^{1,2}$

$^1$Functional Genomics Unit, Institut Jules Bordet

$^2$Machine Learning Group, Université Libre de Bruxelles

October 17, 2008
Research Groups
Machine Learning Group (Gianluca Bontempi)

- 10 researchers (2 Profs, 1 postDoc, 7 PhD students), 2 graduate students).
- Research topics: Bioinformatics, Classification, Regression, Time series prediction, Sensor networks.
- Website: http://www.ulb.ac.be/di/mlg.
- Scientific collaborations in ULB: IRIDIA (Sciences Appliquées), Physiologie Molculaire de la Cellule (IBMM), Conformation des Macromolcules Biologiques et Bioinformatique (IBMM), CENOLI (Sciences), Functional Genomics Unit (Institut Jules Bordet), Service d’Anesthesie (Erasme).
- Scientific collaborations outside ULB: UCL Machine Learning Group (B), Politecnico di Milano (I), Università del Sannio (I), George Mason University (US).
- The MLG is part to the ”Groupe de Contact FNRS” on Machine Learning and to CINBIOS: http://babylone.ulb.ac.be/Joomla/.
9 researchers (1 Prof, 5 postDocs, 3 PhD students), 5 technicians.

Research topics: Genomic analyses, clinical studies and translational research.


National scientific collaborations: ULB, Erasme, ULg, Gembloux, IDDI.

International scientific collaborations: Genome Institute of Singapore, John Radcliffe Hospital, Karolinska Institute and Hospital, MD Anderson Cancer Center, Netherlands Cancer Institute, Swiss Institute for Experimental Cancer Research, NCI/NIH, Gustave-Roussy Institute.
Summary

- Breast Cancer and Prognosis
  - Current Clinical Tools
  - Potential of Genomic Technologies

- Gene Expression Profiling

- Breast Cancer Molecular Subtypes

- Prognostic Gene Signatures

- Subtypes and Prognosis

- Conclusion
Breast cancer is a global public health issue.

It is the most frequently diagnosed malignancy in women in the western world and the commonest cause of cancer death for European and American women.

In Europe, one out of eight to ten women, depending on the country, will develop breast cancer during their lifetime.
Breast Cancer Prognosis

Diagnosis → Breast surgery + radiotherapy → Follow-up 5-10 years → Recurrence → Remission

Prognosis
Need to improve current clinical tools to detect patients who need adjuvant systemic therapy.
In the nineties, new biotechnologies emerged:
- Human genome sequencing.
- Gene expression profiling (low to high-throughput).

Genomic data could be used to better understand cancer biology
... and to build efficient prognostic models.
Part II

Gene Expression Profiling
Biology Paradigm

DNA → Transcription → pre-mRNA → Splicing → mRNA → Translation → protein

Promoter → Enhancers → Introns

Gene expression profiling
There exist several technologies to measure the expression of genes.

Low throughput technologies such as RT-PCR, allow for measuring the expression of a few genes.

High throughput technologies, such as microarrays, allow for measuring simultaneously the expression of thousands of genes (whole genome).

Microarray principles will be illustrated through the Affymetrix technology.
A microarray is based on
- DNA fragments (probes) fixed on a solid support.
- Ordered position of probes.
- Principle of hybridization to a specific probe of complementary sequence.
- Molecular labeling.

Simultaneous detection of thousands of sequences in parallel.
Microarray Chip
Affymetrix Design

1. Total RNA
2. Reverse Transcription
3. In Vitro Transcription
4. Biotin-labeled cRNA
5. Fragmentation
6. GeneChip Expression Array
7. Hybridization
8. Wash and Stain
9. Scan and Quantitate
Part III

Breast Cancer Molecular Subtypes
Breast Cancer Subtypes

- Early microarray studies showed that BC is a molecularly heterogeneous disease [Perou et al., 2000; Sorlie et al., 2001, 2003; Sotiriou et al., 2003].
  - Hierarchical clustering on microarray data [Sorlie et al., 2001]:

  ![Hierarchical clustering diagram](image_url)
Breast Cancer Subtypes
Clinical Outcome

- The molecular subtypes exhibited different clinical outcomes, suggesting that the biological processes involved in patients’ survival might be different.
These early studies showed similar results, i.e. ER and HER2 pathways are the main discriminators in breast cancer (confirmed by [Kapp et al., 2006]).

However, this classification has strong limitations [Pusztai et al., 2006]:

- Instability: the results are hardly reproducible due to the instability of the hierarchical clustering method in combination with microarray data (high feature-to-sample ratio).
- Crispness: hierarchical clustering produces crisp partition of the dataset (*hard partitioning*) without estimation of the classification uncertainty.
- Validation: the hierarchical clustering is hardly applicable to new data.
Because of these limitations we sought to develop a simple method to identify the breast cancer subtypes.

We introduced a model-based clustering in a two-dimensional space defined by the ESR1 and ERBB2 module scores [Wirapati et al., 2008; Desmedt et al., 2008].

- We used the Bayesian information criterion (BIC) to select the most likely number of subtypes.
- We validated our model (fitted on Wang et al. series) on 14 independent datasets.
Breast Cancer Subtypes
Clustering Model

VDX

BIC

number of clusters

ESR1

ERBB2

ER-/HER2-

HER2+

ER+/HER2-

BIC

0

0.2

0.4

0.6

0.8

1

2

4

6

8

10

number of clusters
Breast Cancer Subtypes

Validation

NKI

- ER+/HER2−
- HER2+
- ER+/HER2−

TBG

- ER+/HER2−
- HER2+
- ER+/HER2−

UPP

- ER+/HER2−
- HER2+
- ER+/HER2−

UNT

- ER+/HER2−
- HER2+
- ER+/HER2−

MAINZ

- ER+/HER2−
- HER2+
- ER+/HER2−

UNC2

- ER+/HER2−
- HER2+
- ER+/HER2−

ERBB2

- ESR1
- ERBB2
Breast Cancer Subtypes
Clinical Outcome

- ER-/HER2-: 20-25%
- HER2+: 15-20%
- ER+/HER2-: 60-70% of the global population of BC patients.
Breast Cancer Subtypes
New Clustering Model (dis)Advantages

Advantages:

- Simple model-based clustering:
  - Easily applicable to new data.
  - Returning for each patient the probability to belong to each subtype (soft partitioning).
- Low dimensional space:
  - Low computational cost to fit the model.
  - Simple visualization of the results.

Disadvantages:

- Low dimensional space: which dimension could we add in order to find another robust subtype?
Part IV

Prognostic Gene Signatures
Use of microarray technology to improve current prognostic models (NIH/St Gallen guidelines, NPI, AOL).

A typical microarray analysis dealing with breast cancer prognostication involves 5 key steps:

1. Data preprocessing: quality controls and normalization.
2. Filtering: discard the genes exhibiting low expressions and/or low variance.
3. Identification of a list of prognostic genes (called a gene signature).
4. Building of a prognostic model, i.e. combination of the expression of the genes from the signature in order to predict the clinical outcome of the patients.
5. Validation of the model performance and comparison with current prognostic models.
Prognostic Gene Signatures
Fishing Expedition

- Prognostic models derived from gene expression data by looking for genes associated with clinical outcome without any a priori biological assumption [van't Veer et al., 2002; Wang et al., 2005].

*GENE70 signature*

*GENE76 signature*

- Promising results but a lot criticisms from a statistical point of view.
Prognostic Gene Signatures
Hypothesis-driven

- Prognostic models were also derived from gene expression data based on a biological assumption.
  - Example: GGI [Sotiriou et al., 2006] was designed to discriminate patients with low and high histological grade (proliferation).
  - GGI was able to discriminate patients with intermediate histological grade (HG2).
These preliminary results were promising but validation was required.

A first validation was published by the authors of the GENE70 and GENE76 signatures in [van de Vijver et al., 2002] and [Foekens et al., 2006] respectively.

Our group was involved in a second validation:

- Complete independence: the authors of the signatures were not aware of the clinical data of the patients in the dataset.
- The statistical analyses were performed by an independent group.
- Aim: validate definitively the prognostic power of these two models in order to start a large clinical trial called MINDACT (Microarray In Node negative Disease may Avoid ChemoTherapy).
Although the performance in this validation series was less impressive than in the original publications, GENE70 and GENE76 sufficiently improved the current clinical models to go ahead with MINDACT.

Validation of GENE70 [Buyse et al., 2006] and GENE76 [Desmedt et al., 2007].
We sought to compare the GGI to the GENE70 and GENE76 signatures in this validation series. 

... and showed that GGI has very similar performance [Haibe-Kains et al., 2008b].
From the validation studies, we learned that GGI yields similar (sometimes better) performance than other gene signatures [Haibe-Kains et al., 2008b].

Since GGI is a very simple model from a statistical and a biological (proliferation genes) points of view, we challenged the use of complex statistical methods for BC prognostication.

We compared simple to complex statistical methods to a single proliferation gene (AURKA) [Haibe-Kains et al., 2008a].

Due to the complexity of microarray data, it is very hard to build prognostic models statistically better than AURKA.
Forestplot of the concordance index for each method in the training set and the three validation sets:
Part V

Subtypes and Prognosis
The first publications attempted to build a prognostic model from the global population of BC patients.

In 2005, Wang et al. were the first to divide the global population based on ER status:

- As BC biology is very different according to the ER status, prognostic models might be different too.
- They built a prognostic model for each subgroup of patients (ER+ and ER-).
- To make a prediction, they used one of the two models depending on the ER-status of the tumor.
- Unfortunately the group of ER- tumors was too small and their corresponding model was not generalizable.
Recently, Teschendorff et al. built a new prognostic model for ER-tumors [Teschendorff et al., 2007] and validated it [Teschendorff and Caldas, 2008] using large datasets.
- The signature is composed of 7 immune-related genes.

We showed in two meta-analyses [Wirapati et al., 2008; Desmedt et al., 2008] that:
- Proliferation (AURKA) was the most prognostic factor in ER+/HER2-tumors and the common driving force of the early gene signatures.
  - Actually, these early signatures (e.g. GENE70, GENE76, GGI) are prognostic in ER+/HER2- tumors only.
- Immune response (STAT1) is prognostic in ER-/HER2- and HER2+ tumors.
- Tumor invasion (PLAU or uPA) is prognostic in HER2+ tumors.

Finak et al. introduced a stroma-derived prognostic predictor (SDPP) especially efficient in HER2+ tumors [Finak et al., 2008].
We plan to develop a new prognostic model integrating the breast cancer subtypes identification in order to:

▶ Build prognostic gene signatures targeting a specific subtype.
▶ Build a global prognostic model able to predict the risk of the patients having a tumor of ER-/HER2-, HER2+ or ER+/HER2- subtype.

...and to assess/compare its performance with current prognostic models using the thorough statistical framework developed in [Haibe-Kains et al., 2008a].

We already find a name, GENIUS, standing for Gene Expression progNostic Index Using Subtypes 😊
Part VI

Conclusion
Numerous studies confirmed the great potential of gene expression profiling using microarrays to better understand cancer biology and to improve current prediction models.

This technology becomes more and more mature (MAQC [shi, 2006]) and is now ready for clinical applications.

The promising results of early publications were validated in different independent studies.

Recent meta-analyses successfully recapitulated the main discoveries made these late decades and refined our knowledge on breast cancer biology.

We benefit from this strong basis to go a step further to improve breast cancer prognosis using microarrays, especially by integrating the breast cancer molecular subtypes identification.
Thank you for your attention.


Part VIII

Appendix
Prognosis Using Subtypes

GENIUS: Analysis Design

Subtypes identification

Pr1 Pr2 Pr3

 Identification of prognostic genes
 subtype signature

Model building
 subtype risk score

Combination

Training set

Validation set

Subtype risk score

(Pr1, Pr2, Pr3)

Pr1 Pr2

Identification of prognostic genes
 subtype signature

Model building
 subtype risk score

Genius

GNI

Identification of prognostic genes
 subtype signature

Model building
 subtype risk score

Combination

GENIUS

Risk score

Risk group

Performance assessment and comparison

Validation set

Training set

ER-/HER2- HER2+ ER+/HER2-

Subtypes identification

Pr1 Pr3 Pr2

Identification of prognostic genes
 subtype signature

Model building
 subtype risk score

Combination

GENIUS

Risk score

Risk group

Performance assessment and comparison
Bioinformatics softwares

- **R** is a widely used open source language and environment for statistical computing and graphics
- **Bioconductor** is an open source and open development software project for the analysis and comprehension of genomic data
- **Java Treeview** is an open source software for clustering visualization
- **BRB Array Tools** is a software suite for microarray analysis working as an Excel macro
Links

- Personal webpage: http://www.ulb.ac.be/di/map/bhaibeka/
- Master in Bioinformatics at ULB and other belgian universities: http://www.bioinfomaster.ulb.ac.be/