A bioinformatics pipeline for microarray analysis: from cell models to breast cancer classification

Isella C, Renzulli T, Medico E

Institute for Cancer Research and Treatment,
University of Torino Medical School

Motivation

Breast cancer accounts for ~30% of all cancers diagnosed in women and for ~16% of cancer deaths, due to metastatic progression. However, the term breast cancer refers to a wide spectrum of diseases, with a large range of clinical outcomes. Indeed, current clinico-pathological parameters fail to correctly predict relapse in a significant fraction of cases; as a consequence, adjuvant chemotherapy is given to many patients who will not take advantage from it. Over the last few years, DNA microarrays have been employed to measure expression of thousands of genes in cancer samples from patients with different outcome, leading to the empirical design of genomic predictors of metastatic propensity. However these classifiers were validated on restricted independent validation sets, and, being derived from several thousands of variables, are affected by overfitting. Here we propose an automated pipeline to evaluate predetermined gene lists, derived from genomic explorations in experimental models, for their ability to discriminate breast cancer outcome by constructing genomic classifiers derived from these lists and validating the results on a wide independent dataset of breast cancer patient.

Methods

Taking advantage of the large amount of published gene expression profiles for breast cancer, we merged several datasets encompassing a total of over 1300 samples. Array probes were univocally cross-mapped to single genes using a “MAQC” table. Gene sets of interest underwent a multistep pipeline, implemented in R-Bioconductor, in which:
i) Gene lists are mapped to human breast cancer microarray datasets.
ii) A signature is evaluated for its enrichment in genes discriminating poor prognosis patients from good prognosis patients.
iii) In the training subset, a genomic classifier is build according to a Nearest Mean Classifier model (NMC).
iv) The genomic classifier is validated by ROC and log-rank analysis on an independent dataset.
v) Montecarlo simulations are carried out to evaluate whether the obtained performance can be reached also starting from a randomly selected gene list of the same size.

Results

We applied the pipeline to gene sets generated by transcriptome analysis of in vitro biological models related to cancer progression. Namely a list of 83 genes differentially expressed in MCF10A normal breast cells rendered anchorage-independent by GAB2 overexpression, a list of 27 genes differentially expressed in MLP-29 mouse embryo liver cells in which invasive growth is activated in response to HGF, a list of 228 genes that are targets of miR-24 and regulated by HGF, and a control list of genes with no evident relation with breast cancer progression. Interestingly, a well-performing classifier was obtained and validated starting from many random gene sets, but only the biological model-derived sets showed a significantly higher discrimination power compared to their respective Montecarlo simulations. This work shows that many randomly selected gene lists can achieve interesting classifying performances, but only genomic signatures derived from in-vitro models related to cancer progression can construct outstanding classifiers. This strategy can be applied to any set of genes for evaluating its collective ability to predict breast cancer outcome beyond a random probability.

Contact : enzo.medico@ircc.it