Chapter 8
Linking Interactome to Disease:
A Network-Based Analysis of Metastatic Relapse in Breast Cancer

Maxime Garcia
Inserm, Paoli Calmettes Institute, France

Daniel Birnbaum
Inserm, Paoli Calmettes Institute, France

Olivier Stahl
Inserm, Paoli Calmettes Institute, France

François Bertucci
Inserm, Paoli Calmettes Institute, France

Pascal Finetti
Inserm, Paoli Calmettes Institute, France

Ghislain Bidaut
Inserm, Paoli Calmettes Institute, France

ABSTRACT

The introduction of high-throughput gene expression profiling technologies (DNA microarrays) in molecular biology and their expected applications to the clinic have allowed the design of predictive signatures linked to a particular clinical condition or patient outcome in a given clinical setting. However, it has been shown that such signatures are prone to several problems: (i) they are heavily unstable and linked to the set of patients chosen for training; (ii) data topology is problematic with regard to the data dimensionality (too many variables for too few samples); (iii) diseases such as cancer are provoked by subtle misregulations which cannot be readily detected by current analysis methods. To find a predictive signature generalizable for multiple datasets, a strategy of superimposition of a large scale of protein-protein interaction data (human interactome) was devised over several gene expression datasets (a total of 2,464 breast cancer tumors were integrated), to find discriminative regions in the interactome (subnetworks) predicting metastatic relapse in breast cancer. This method, Interactome-Transcriptome Integration (ITI), was applied to several breast cancer DNA microarray datasets and allowed the extraction of a signature constituted by 119 subnetworks. All subnetworks have been stored in a relational database and linked to Gene Ontology and NCBI EntrezGene annotation databases for analysis. Exploration of annotations has shown that this set of subnetworks reflects several biological processes linked to cancer and is a good candidate for establishing a network-based signature for prediction of metastatic relapse in breast cancer.

DOI: 10.4018/978-1-4666-3604-0.ch008

INTRODUCTION

Since introduction of high-throughput technologies in molecular biology in the late nineties, a number of technologies for deciphering the genomic origin of several diseases has flourished. Among these, cDNA microarrays (Schena et al., 1995) have allowed measuring Gene Expression Profiles (GEP) at the genome scale and have shed light on large scale gene regulation/misregulation under varied conditions. Many diseases, including several forms of cancer [leukemia (Golub et al., 1999), colon cancer (Li et al., 2001), breast cancer (Wang et al., 2005)], diabetes (Kaestner et al., 2003), and others (Munro & Perreau. 2009) have been studied that way. Of particular interest in the context of cancer, a particularly heterogeneous disease, is the use of GEPs to either predict drug resistance (de Lavallade et al, 2010), or the metastatic recurrence, for instance in breast cancer (van de Vijver et al., 2002). Tumor microenvironment studies have allowed understanding the influence of immune system on patient outcome (Pagès et al., 2009).

There is an increasing number of controversial cases for the use of systemic adjuvant therapy due to the clinical and pathological heterogeneity of the disease to treat. In node-negative early breast cancer, most patients undergo adjuvant chemotherapy even though 70-80% of them would have survived without it (Bertucci & Birnbaum, 2008). The refinement of current prognostic histopathological methods using molecular diagnostics can also lead to increase of detection of disease subtypes that necessitate specific treatments, such as T1 breast cancer (Mook et al., 2010). In all cases, the goal is to refine and individualize treatment and lead the way to personalized medicine for a growing number of pathologies.

In cancer, the understanding of molecular basis of metastasis is of primary importance. Several studies have attempted to obtain a molecular portrait for a large number of patients using DNA microarray analysis, performed supervised analysis and published list of genes predicting patient outcome. Two of these signatures are currently under clinical trials in breast cancer: the MINDACT trial, based on the 70-genes Mammaprint signature [van't Veer et al. (2002) van de Vijver et al. (2002), Bueno-de-Mesquita et al. (2007)]), and the TAYLORx trial, an RT-PCR-based 21-genes OncotypeDX signature (Paik et al., 2006).

However, most prognostic signatures reported for breast cancer show very little or no overlap, and do not appear generalizable from one study to another, and this un-reproducibility was widely criticized (Chuang et al., 2007, Bertucci et al., 2008). Two studies in particular are often cited for their lack of agreement, although they addressed similar questions, which were the two breast cancer studies performed by van de Vijver et al. (2002) and Wang et al. (2005), who reported two prognosis signatures for metastatic relapse in breast cancer. Two different signatures comprising respectively 70 and 76 genes predictive of breast cancer patient outcome were reported but presented only three genes in common. Even more concerning was the study by Michiels et al. (2005) which showed that hundreds of 70-genes signatures with equal classification power can be drawn by shuffling the training and test sets in the van't Veer et al. (2002) dataset, showing the instability and dependency of the resulting gene lists on the training data.

Microarray technology itself was blamed at first for these inconsistencies, and DNA microarrays were suspected to be extremely noisy and leading to non reproducible results. Once the stability and inherent reproducibility were demonstrated by comparing several platforms in the Microarray Quality Control project (Shi et al., 2006), the reasons for the lack of uniqueness in gene signatures had to be found elsewhere.

This chapter deals with addressing the problem of signature instability and proposes a new computational model, the Interactome-Transcriptome Integration (ITI), which simultaneously integrates multiple datasets, to compensate for data dimensionality, and uses the human interactome to include genes with weaker signal in the signature.