Breast Cancer, Microarrays and Biomedical Informatics: 
The Prognochip Project
(extended abstract)

G. Potamias\textsuperscript{1,2}, A. Analyti\textsuperscript{1}, D. Kafetzopoulos\textsuperscript{3}, M. Tsiknakis\textsuperscript{1}, D. Plexousakis\textsuperscript{1,2},
P. Poirazi\textsuperscript{3}, M. Reczko\textsuperscript{1}, Y. Tollis\textsuperscript{1,2}, E. Sanidas\textsuperscript{3}, E. Stathopoulos\textsuperscript{5}, S. Vassilaros\textsuperscript{6}

\textsuperscript{1}Institute of Computer Science (ICS), FORTH, \textsuperscript{2}Dept. of Computer Science, University of Crete, \textsuperscript{3}Institute of Molecular Biology and Biotechnology (IMBB), FORTH, \textsuperscript{4}Dept. of Surgical Oncology, Heraklion Medical School, Crete, Greece, \textsuperscript{5}Dept. of Pathology, Heraklion Medical School, Crete, Greece, \textsuperscript{6}Prolipsis Diagnostic Breast Center, Athens, Greece

1. Introduction

Breast cancer is one of the most common malignancies affecting women, the lifetime risk being approximately 10%. Breast cancer (BRCA) is both genetically and histopathologically heterogeneous, and the mechanisms underlying BRCA development remains largely unknown. Although, conventional prognostic indicators such as lymph node status, oestrogen receptor (ER) status or histological grade are extremely valuable, it is still particularly difficult to predict which patients will develop metastases.

Global expression analysis using microarrays now offers unprecedented opportunities to obtain molecular signatures of the state of activity of diseased cells and patient samples. This groundbreaking approach of studying cancer promises to provide a better understanding of the underlying mechanism for tumorigenesis, more accurate diagnosis, more comprehensive prognosis, and more effective therapeutic interventions. Recent BRCA studies have demonstrated the ability of microarray-based expression profiling to detect tumor cells in peripheral blood samples, to predict chemotherapy responses in fine-needle aspiration samples in neoadjuvant chemotherapy, and, most importantly, to predict disease-free survival and overall survival from profiles in breast cancer surgical specimens. The predictive power of this approach is much greater than that of currently used approaches, but remains to be validated in prospective clinical studies.

The goal is quite challenging with a major prerequisite namely, the cooperation between the involved scientific disciplines. It becomes evident that in order to fully grasp the mechanisms of a disease we do not only need an understanding of the genetic base of the disease-dealing with large amounts of data and related functional genomics approaches (such as gene-expression profiling) but we also need to integrate the knowledge normally processed in the clinical setting. In other words the R&D agenda should be forwarded towards the integration or, synergy between bioinformatics and medical informatics activities.

In this setting a new discipline namely, Biomedical Informatics (BMI), is raising. BMI aims to offer the appropriate technology in order to support the emerging ‘individualised medicine’ environment, and allow optimised, individualized healthcare using all relevant sources of information. The use of genetic and proteomic data in addition to clinical symptoms for medical decision-making will contribute to the expected, continued shift towards evidence-based medicine. This vision can only be realized with an enormous investment into: (i) technology able to produce the genomic and proteomic data and the initial comparison of produced results with reference databases; (ii) creation of standardized databases that combine clinical history, symptoms and signs, laboratory and procedural results, and genetic and proteomic data in raw as well as intelligently processed formats; (iii) technology that assures confidential access to these data by those who need access, and full-proof security against unauthorized access; (iv) extraction of knowledge out of these huge databases, their expert interpretation and matching against existing computational models; (vi) development of novel explanatory and predictive models for the above, abstraction of the results to the clinical level, and incorporation of the extracted knowledge into algorithms and standardized clinical guidelines where feasible; and finally (vii) implementation of the new guidelines into the clinical decision-making process.

In the next sections we present the key-ideas underlying the Prognochip project (funded by the Greek General Secretariat for Research & Technology in the context of the EPAN programme, 2003-2006) where, the aforementioned objectives and goals seek their realization.
2. Prognochip: Microarrays for the Identification and Validation BRCA Molecular Markers

*Prognochip* is a (running) project that joins forces and efforts from different scientific disciplines: Molecular Biology [Institute of Molecular Biology & Biotechnology, FORTH], Medicine [University Hospital, University of Crete, and PROLIPSIS, diagnostic breast cancer center], and Computer Science [Institute of Computer Science, FORTH]. The major tasks (already scheduled and initiated) within Prognochip are briefly presented into the sequel.

**Ethical Issues:** Patients are to be informed and consent to the molecular analysis of their tissue and blood samples; they will also consent for the use of the study’s data for scientific reasons provided that their anonymity is secured; a respective ‘consent form’ is already under the approval of the ‘Ethics and Scientific Committee’ of the University Hospital at Heraklion, and the PROLIPSIS breast cancer center.

**Molecular Biology/ Microarrays:** (a) Design of the BRCA microarray chip - selection of BRCA related or specific genes; (b) Choice of positive & negative control and spike-in controls that are required for the evaluation of the sample preparation & hybridization processes as well as for the statistical processing of the measurements. (c) Manufacturing of the BRCA microarray chips - the related oligos will be deposited on coated slides in the humidity-conditioned cabinet of a Packard SpotArray 24 spotter; after spotting a set of quality controls will be routinely performed using dyes & standard samples on a GSI Lumonics ScanArray 5000 scanner; the aforementioned methods have been well established in the post-genomics facility of IMBB, being fully equipped and operational (http://www.imbb.forth.gr/facilities/genomic.html); (d) Gene expression profiling from ~300 patients is planned throughout the project; the experimental design involves the choice of standards for the comparison with the diseased tissues and the avoidance of redundancies among the tissues to be analyzed; the following, protocoled steps, are provisioned: (i) tissue RNA extraction – to be performed in such a way that verifies the stabilization of the RNA; (ii) labelling - using unbiased chemical bonding of fluorescent groups with two different emission wavelengths to the nucleic acids; (iii) hybridization to microarrays - in order to identify and form different BRCA sub-classes and discover potential diagnostic and prognostic molecular markers; (iv) fluorescence scanning - on a GSI Lumonics ScanArray 5000 scanner the fluorescence intensities will be translated in 16-bit TIFF files, one for each wavelength; and (v) Image analysis - using suitable software (e.g., QuantArray already purchased by IMBB), each image will be analyzed and transformed in a spreadsheet full of valuable data on the individual spot intensities.

3. Biomedical Informatics: Towards an Integrated Clinico-Genomics Environment

Collaborative efforts between Medical Informatics (MI) and Bioinformatics (BI) could provide new insights and create a synergy for challenges needed to create novel genomic applications in medicine (refer to http://bioinformed.isciii.es/). BI enables us to understand the fundamental knowledge about biological processes. The inclusion of clinical information in biomedical informatics opens the gateway to genetic risk profiling of patients, new paradigms in disease diagnoses and prognoses and novel approaches to drug discovery based on the correlation of genetic and molecular knowledge of diseases with clinical information of the patients. In this setting the respective biomedical informatics R&D agenda is forwarded towards the design, development and deployment of an integrated clinico-genomics operational framework where, functional genomics and disease compacting research are coupled and guided by related medical knowledge. The endeavour is to be based on the synergy between MI and BI and centred on the promising microarray technology.

In the context of the Prognochip we have forwarded, scheduled and initiated efforts towards the delivery of an *Integrated Clinico-Genomics Information Technology Environment* (ICG-ITE) with the combined genetic- and individualized-medicine being the target. The envisioned building blocks of ICG-
ITE include: (a) a set of clinical information systems to keep patients’ clinical information (i.e., clinical, laboratory and histo-patholo-anatomy information systems); (b) an information system to store and manage the specifications of the respective microarray experiments (i.e., chip design, hybridizations, etc), analyze measured biossays, as well as to store samples’ genomic information (GIS: Genomic Information System), and (c) a middleware layer for information/data integration and intelligent processing - realized by a ‘puzzle’ of integrated software components that enable: (i) the seamless and efficient extraction of data from the various data and information sources (clinical and genomic); (ii) uniform information modeling - enabled by the utilization of standard clinical genomics data models and respective ontologies; (iii) uniform information representation - enabled by the utilization and the appropriate customization of RDF/XML technology; and (iv) intelligent data processing and visualization - enabled by a suite of data-mining components and tools. The demanding clinical and genomic data integration environment post the need to elaborate on the concept of Integrated Electronic Health Care Record (IEHCR) architectures, utilize the respective technological advances, and extend the standard clinical data models to include and amalgamate genomic ones. In this context, the provided security and authorisation infrastructure is fully employed. A general layout of the provisioned ICG-ITE is shown in Figure 1.

Figure 1. Building Blocks, Components and Operations in the Integrated Clinico-Genomics Information Technology Environment (ICG-ITE)