Artificial Intelligence in Medicine 53 (2011) 57–71

Contents lists available at ScienceDirect

Artificial Intelligence in Medicine

journal homepage: www.elsevier.com/locate/aiim

Integration of gene signatures using biological knowledge

Michalis E. Blazadonakis[a,*], Michalis E. Zervakis[a], Dimitrios Kafetzopoulos[b]

[a] Department of Electronic and Computer Engineering, University Campus, Technical University of Crete, GR 731 00, Chania Crete, Greece
[b] Institute of Molecular Biology and Biotechnology, Foundation for Research and Technology - Hellas, N. Plastira 100, GR 700 13, Heraklion Crete, Greece

ARTICLE INFO

Article history:
Received 1 November 2009
Received in revised form 14 May 2011
Accepted 16 June 2011

Keywords:
Marker gene selection
Gene signature integration
Cross-platform integration
Biological knowledge
Gene ontology
Pathways
Breast cancer

ABSTRACT

Objective: Gene expression patterns that distinguish clinically significant disease subclasses may not only play a prominent role in diagnosis, but also lead to the therapeutic strategies tailoring the treatment to the particular biology of each disease. Nevertheless, gene expression signatures derived through statistical feature-extraction procedures on population datasets have received rightful criticism, since they share few genes in common, even when derived from the same dataset. We focus on knowledge complementarities conveyed by two or more gene-expression signatures by means of embedded biological processes and pathways, which alternatively form a meta-knowledge platform of analysis towards a more global, robust and powerful solution.
Methods: The main contribution of this work is the introduction and study of an approach for integrating different gene signatures based on the underlying biological knowledge, in an attempt to derive a unified global solution. It is further recognized that one group's signature does not perform well on another group's data, due to incompatibilities of microarray technologies and the experimental design. We assess this cross-platform aspect, showing that a unified solution derived on the basis of both statistical and biological validation may also help in overcoming such inconsistencies.
Results: Based on the proposed approach we derived a unified 69-gene signature, which outperforms significantly the performance of the initial signatures succeeding a 0.73 accuracy metric on 234 new patients with 81% sensitivity and 64% specificity. The same signature manages to reveal the two prognostic groups on an additional dataset of 286 new patients obtained through a different experimental protocol and microarray platform. Furthermore, it manages to derive two clusters in a dataset from a different platform, showing remarkable difference on both gene-expression and survival-prediction levels.

1. Introduction

Microarray technology has become a valuable tool for classifying breast tumors according to their prognosis, subtype, or response to treatment. An open problem in studies of breast cancer, as well as other types of cancer, is to determine the most appropriate treatment protocol for a specific patient. Even though chemotherapy or hormonal therapy reduces the risk of distant metastasis by approximately 1/3, 70–80% of the patients receiving adjuvant treatment would have survived without it [1,2]. Along with the treatment plan, there are also variations in the evaluation of diagnostic means. Existing inconsistencies in histological grading forced the American Joint Committee on Cancer to exclude histological tumor grading from its staging criteria [3]. Hence, attempts to increase the prognostic value through the use of stable and robust markers become more than a necessity and this is a direction towards which microarray technology is expected to contribute.

The mass information conveyed by microarray data must be robustly processed in order to enable the extraction of meaningful and reproducible results necessitating close collaboration among many scientific fields, such as medicine, biology, statistics and computer science. Besides the need of statistical validation, "an understanding of both the biology and the computational methods is essential for tackling the associated data mining task without being distracted by the abundant fool's gold" [4]. We express these issues by adopting the position that simply generating statistically significant results is not enough in genomic analysis; any result should also be evaluated in terms of its biological significance, which in any case is clinically more important than its statistical relevance on a limited dataset. Nevertheless, since the recorded knowledge may not be complete, the biological significance is used in our approach to enrich the statistical result rather than to filter it out of potentially noisy and biologically irrelevant genes. To further support this position we refer to the study of Van't Veer et al. [5], which has received criticism [6,7] from a statistical point of view when considering stringent statistical criteria. Neverthe-

* Corresponding author. Tel.: +30 2821037206; fax: +30 2821037542.
E-mail addresses: mblazad@gmail.com (M.E. Blazadonakis), michalis@display.tuc.gr (M.E. Zervakis), kafetzo@imbb.forth.gr (D. Kafetzopoulos).

0933-3657/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.artmed.2011.06.003

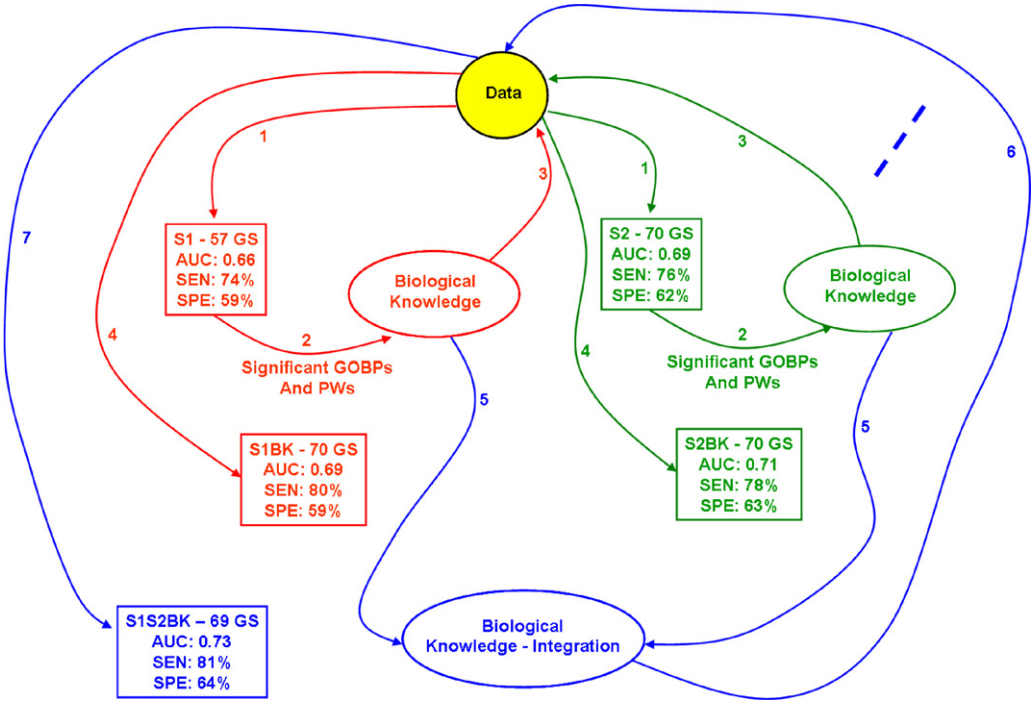

Fig. 1. The proposed biological-knowledge integration process that derives the S1S2-BK gene signature; red-green part represents independent focus on a single signature; blue part represents biological knowledge integration. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

less, taking into account additional biological and medical criteria, the Food and Drug Administration (FDA) [8] approved the result of Van't Veer et al. for further clinical tests. Specifically, the 70-gene derived by the study has been approved by FDA under the product name 'MammaPrint' [9] as the first clear product that profiles gene expression correlated to the likelihood of tumor recurrence. Furthermore, the European Organization for Research and Treatment of Cancer (EORTC) [10] launched the MINDACT [11] (microarray in node negative disease may avoid chemotherapy) project opening the road for the evaluation of gene expression signatures in randomized clinical trials.

In this study we demonstrate that, by appropriately adopting biological knowledge, the statistical results can be significantly improved. Published gene signatures have few or perhaps no genes in common. Recent studies, however, prove that even if there is no significant gene overlap between two different gene signatures, there might be significant overlap in terms of the biology conveyed by them [12]. Building on these steps, we search for and exploit significant biological information by means of gene ontology biological processes (GOBPs) and pathways hidden behind a 57-gene signature (referred to as S1), the derivation of which is discussed extensively in [13]. In parallel we consider one widely discussed signature, i.e. the 70-gene signature (referred to as S2) published by Van't Veer et al. in [5]. First, we show that the two signatures indeed demonstrate significant biological overlap when GOBPs and pathways are considered, in complete agreement to [12]. Secondly, we demonstrate that statistical results are substantially improved when integrating the biological knowledge conveyed by the individual signatures S1 and S2, which have been derived from the same dataset. The result of such knowledge integration is evaluated on the 234 new cases published in [15], as well as on the 286 cases published in [14] as to assess its efficiency on a different microarray platform and experimental design. The proposed approach actually addresses two major problems in cancer gene-selection: (a) the issue of minor or no overlap between different gene signatures at the gene level [16], and (b) the cross platform evaluation of results by testing a predictor that is derived using a specific microarray platform and experimental design on another group's data derived with a different experimental platform and protocol [18].

As stated in the abstract, the main contribution and novelty of this work is that it proposes an approach for integrating different gene signatures, in an attempt to derive a unified and more global solution based on the common biological aspects of the individual signatures. Even though several integration schemes have been proposed in the area of microarray analysis (selectively refer to [12,19–21]), little has been done towards assessing the concept of integrating results from different group. Our motive behind this approach is the meta-analysis view that different gene signatures, which may be proposed by different research groups, form the parts of a more global solution with each individual approach addressing only a small part of the whole. In turn, we treat such individual (and seemingly different) solutions as pools of valuable knowledge for integration and unification of solutions. We emphasize that this process is a dynamic one, which can systematically evolve to include more individual gene-signatures related to the same pathology.

The concept of gene-signature overlap can be considered at four levels of abstraction as follows:

1. Gene-level overlap assesses the number of common genes between two signatures. This issue has been addressed before showing minimal or no overlap among the various signatures published in breast cancer [16,18].
2. Pathway-level overlap assesses the common pathways that exist between two gene signatures [16,18]. We should clarify here that the term "pathway" encompasses all processes associated with both GOBPs and pathways.
3. Overlap of significant pathways is measured by HGPD discussed in Section 2. After specifying all pathways induced by the genes of a gene signature, we can use the value of HGPD in order to produce a rank-order list of significance. We then define significant