Mathematical Biosciences 260 (2015) 25–34

Contents lists available at ScienceDirect

Mathematical Biosciences

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

Pathway-level disease data mining through hyper-box principles

Lingjian Yang [a,1], Chrysanthi Ainali [b,c,1], Aristotelis Kittas [b], Frank O. Nestle [c], Lazaros G. Papageorgiou [a,*], Sophia Tsoka [b,*]

[a] Centre for Process Systems Engineering, Department of Chemical Engineering, University College London, Torrington Place, London WC1E 7JE, UK
[b] Department of Informatics, School of Natural and Mathematical Sciences, King's College London, Strand, London WC2R 2LS, UK
[c] St John's Institute of Dermatology, Division of Genetics and Molecular Medicine, King's College London School of Medicine, Tower Wing, Guy's Hospital, Great Maze Pond, London SE1 9RT, UK

ARTICLE INFO

Article history:
Received 12 May 2014
Revised 11 September 2014
Accepted 13 September 2014
Available online 19 September 2014

Keywords:
Disease classification
Pathway-based classification
Mathematical programming
Hyper-box-representation
Mixed integer optimisation

ABSTRACT

In microarray data analysis, traditional methods that focus on single genes are increasingly replaced by methods that analyse functional units corresponding to biochemical pathways, as these are considered to offer more insight into gene expression and disease associations. However, the development of robust pipelines to relate genotypic functional modules to disease phenotypes through known molecular interactions is still at its early stages.

In this article we first discuss methodologies that employ groups of genes in disease classification tasks that aim to link gene expression patterns with disease outcome. Then we present a pathway-based approach for disease classification through a mathematical programming model based on hyper-box principles. Association rules derived from the model are extracted and discussed with respect to pathway-specific molecular patterns related to the disease. Overall, we argue that the use of gene sets corresponding to disease-relevant pathways is a promising route to uncover expression-to-phenotype relations in disease classification and we illustrate the potential of hyper-box classification in assessing the predictive power of functional pathways and uncover the effect of specific genes in the prediction of disease phenotypes.

1. Prediction of disease outcome from gene expression

In complex diseases the current diagnostic and prognostic factors fail to accurately distinguish patients of different clinical outcomes [1–7], mainly because of the heterogeneous nature of disease aetiology and poorly understood underlying mechanisms [8]. The advent of high throughput methods, where snapshots of gene or gene product activity are profiled in samples of varying disease states, has provided a powerful means of relating disease phenotypes to latent genetic mechanisms [9]. However, in the use of gene expression measurements for discovering which gene biomarkers (out of thousands) can best predict disease outcome, the relevant classification task suffers from the curse of dimensionality, where learning from a finite number of data samples in a high dimensional feature space requires a large number of training samples. Therefore, to obtain statistically reliable gene signatures capable of accurate prediction of samples into their corresponding phenotypic outcomes, suitable reduction of dimensionality is sought [10,11].

Dimension reduction aims to reduce the number of features in the original gene expression matrix to a size that is conducive for efficient disease classification [12–15]. Traditional computational methods estimate the discriminative power of individual genes, whose expression pattern can best distinguish the samples, and a classifier is subsequently applied on the reduced set of differentially expressed genes. However, gene-based disease classification approaches have been shown in various recent studies to yield non-reproducible gene signatures [16–18]. Thus, current studies have increasingly focused on integrating biological data a priori, including protein interaction networks (PIN) [18] or biological pathway information [19] into gene expression data and proposing biomarkers at the level of functional sets of genes. It has been demonstrated that biomarkers represented as *groups* of genes result in higher prediction accuracy, more reproducibility across different datasets and better mechanistic interpretation [18, 20–23].

1.1. Single gene-based approaches

In general, disease classification approaches where all genes are considered simultaneously employ a feature selection algorithm

* Corresponding authors:
E-mail addresses: l.papageorgiou@ucl.ac.uk (L.G. Papageorgiou), sophia.tsoka@kcl.ac.uk (S. Tsoka).
[1] These authors contributed equally.

http://dx.doi.org/10.1016/j.mbs.2014.09.005
0025-5564/

to select a subset of highly differentially expressed genes from the entire set of profiled genes, before training a classifier on the selected genes. In [4,24], genes were rank-ordered by their degree of correlation with the breast cancer distant-metastasis-free survival time and breast cancer outcome respectively, and gene signatures have been constructed by sequentially adding genes from the ranked list until the maximum prediction performance is reached. A similar sequential gene selection method has been described in [25], which takes into account both the discriminative power of individual genes and their correlation. The underlying idea is that addition of a new candidate gene into the optimal gene set must contribute towards increased classification performance while maintaining a low level of correlation with the current genes in the set to reduce redundant information. An optimisation model that outputs a user-specified subset of genes is proposed in [26], where group of genes is selected to achieve maximal cross-class separability, minimal same-class tightness and gene pair-wise correlation. Similar methods are proposed in other studies [27–31].

A number of ensemble methods are also available, which combine the advantages of many different classification methods that are used in tandem, so as to derive a more efficient framework in comparison to stand-alone feature reduction and single classifiers. For example, principal component analysis is employed in [32] to project the expressions of genes into 10 dominant principal components, followed by inner cross validation procedure where an artificial neural network classifier was been trained for each training sample subset. All constructed classifiers subsequently cast a vote to determine the phenotype of a testing sample. In [33] a list of genes with best discriminative power is constructed and then assigned to different gene subsets. A neural network classifier is trained with each subset of genes and the final ensemble classifier is formed with majority voting strategy. Given a training sample set, [34] creates a number of bootstrap sample sets by drawing with replacement and determining for each bootstrap sample set the weights of genes. Ensemble feature ranking is achieved by aggregating gene weights over all bootstrap sample sets to produce a final gene ranking. In [35] information gain is used to evaluate both the separability of genes and gene–gene dependence, followed by clustering genes into different gene groups with a Markov blanket. Subsequently, different gene sets are constructed by randomly sampling one representative gene from each gene group, and an ensemble classifier is built by learning from each gene set. Similar ensemble-based approaches using microarray samples for disease classification are reported in [36–39].

Gene markers constructed across different datasets share disappointingly little overlap, despite similar predictive power [16–18]. This lack of agreement limits the applicability of such methods and renders mechanistic interpretability problematic. Complex diseases are the consequence of perturbations of genes that act concurrently, rather than genes that act in isolation [40]. Therefore, research efforts have increasingly focused on integrating microarray gene expression profiles with protein interaction network or expert-curated biological pathways, to yield biomarkers through functional sets of genes [16,19,41], discussed next.

1.2. Network-based approaches

The past few years have seen remarkable growth of protein interaction data. Network principles, where nodes represent genes or their products and edges indicate some type of interaction between them, have served as a particularly suitable abstraction bases to develop computational procedures for understanding system properties [40]. Disease module-based methods assume that all cellular components that belong to the same topological, functional or disease module have a high likelihood of being involved in the same disease [42,43]. This strategy involves constructing the interactome by integrating available data from online databases in the tissue or cell line of interest and then identifying functional units that contain most of the disease-associated genes. Disease modules are then validated by, for example, showing that the genes in a module have related functions or correlated expression patterns [40].

Several methods have been proposed that extract functional modules of genes, whose expression patterns can distinguish samples from different phenotypes [18,20, 44-46]. Analysis of two breast cancer patient cohorts have revealed that altered modularity of the human interactome may be useful as an indicator of breast cancer prognosis [47]. In [18,46] a greedy search is performed over a PIN network to identify a number of gene modules whose average expression is locally maximal. The averaged expression values per module for each sample, called *module activity*, are used as features for a subsequent classification task. Discriminative power and correlation among genes in a linear path search is employed [20]. Linear paths are then combined to form modules, and module activity is inferred with a probabilistic method. In [48], a traditional support vector machine (SVM) classifier is modified to consider the structure of gene interactions and force adjacent genes to contribute similarly when building the classifier. Minimal modules where numbers of differentially expressed member genes exceeded a pre-specified threshold are investigated in [49].

Each of those disease modules may reflect a specific functional pathway relevant to development of the disease of interest. However, it is argued that PIN data is generally unreliable and noisy, as PIN networks typically represent a collection of many interactions under various experimental conditions and cell cycle phases. Another problem of using PIN data is the high false positive rate, meaning that part of recorded interactions may not actually exist [50,51], which in turn may lead to false discovery of biomarkers. Therefore, the adoption of canonical pathways, rather than protein interaction modules, using expertly curated knowledge may provide clearer insight into the interplay between genes in complex diseases.

1.3. Pathway-based approaches

Biological pathways are a trusted expert-curated collection of molecular interaction networks. The increasing availability of pathway information from databases, such as Reactome [52] and the Kyoto Encyclopedia of Genes and Genomes (KEGG) [53], along with annotation databases such gene ontology (GO) [54], is making pathway-based gene set evaluation increasingly popular. Traditional pathway analysis methods include gene set enrichment analysis (GSEA) [55], which produces a list of genes according to their correlation with class label, and calculates for each gene set an enrichment score quantifying its deviation from random distribution, where higher values indicate increased likelihood of the gene set to be disease-relevant. An extension to the original gene set enrichment is suggested in [56] by replacing the two-sample t-test with a linear regression model to assess the gene-phenotype dependence, while in [57] the enrichment analysis is formulated as testing the hypothesis that mean vectors of expression patterns of member genes in a set do not differ in different phenotypes.

In disease classification studies, the concept of using sets of genes that reflect specific biochemical pathways has been proposed as a promising alternative to account for the joint effects of genes involved in similar functional units and, therefore, derive stronger associations with disease phenotype [58,59]. Such approaches have the potential to address genetic dissimilarities across patients and cellular heterogeneity within tissues that yield weak associations among genes and decreased classification performance [19,60]. Here, we present a pathway level disease classification approach based on hyper-box principles, where given a microarray gene expression profile and a number of biological pathways/gene sets, we evaluate the classification accuracy for each pathway using only the member genes in