Computational Biology and Chemistry xxx (2015) xxx–xxx

Contents lists available at ScienceDirect

Computational Biology and Chemistry

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

Research Article

A new vision of evaluating gene expression signatures

Hung-Ming Lai[a,*], Celal Özturk[b], Andreas Albrecht[c], Kathleen Steinhöfel[a]

[a] Algorithms and Bioinformatics Research Group, Department of Informatics, King's College London, Strand, London WC2R 2LS, UK
[b] Department of Computer Engineering, Faculty of Engineering, Erciyes University, Kayseri 38039, Turkey
[c] School of Science and Technology, Middlesex University, Burroughs, London NW4 4BT, UK

ARTICLE INFO

Article history:
Received 2 January 2015
Accepted 3 February 2015
Available online xxx

Keywords:
Bead-chain plot
Gene signature evaluation
Near-optimal signature
Phenotype distinction
Swarm intelligence

ABSTRACT

Gene expression profiles based on high-throughput technologies contribute to molecular classifications of different cell lines and consequently to clinical diagnostic tests for cancer types and other diseases. Statistical techniques and dimension reduction methods have been devised for identifying minimal gene subset with maximal discriminative power. For sets of *in silico* candidate genes, assuming a unique gene signature or performing a parsimonious signature evaluation seems to be too restrictive in the context of *in vitro* signature validation. This is mainly due to the high complexity of largely correlated expression measurements and the existence of various oncogenic pathways. Consequently, it might be more advantageous to identify and evaluate multiple gene signatures with a similar good predictive power, which are referred to as near-optimal signatures, to be made available for biological validation. For this purpose we propose the bead-chain-plot approach originating from swarm intelligence techniques, and a small scale computational experiment is conducted in order to convey our vision. We simulate the acquisition of candidate genes by using a small pool of differentially expressed genes derived from microarray-based CNS tumour data. The application of the bead-chain-plot provides experimental evidence for improved classifications by using near-optimal signatures in validation procedures.

1. Background

The analysis of gene expression data plays an important role in biomedical research. High-throughput screening (HTS), such as microarray data, is still the method of choice when it comes to obtaining gene expression profiles for cells of interest. The analysis of the obtained expression profiling data has shown enormous potential for the discovery of biomarkers in carcinogenesis studies and in the diagnoses of diseases in general (Nevins and Potti, 2007). Different types of tumour cells can be marked by discriminating genes at expression level. Thus, biomarkers for distinct tumourigenesis stages and cancer classification can be identified by selecting discriminating genes, a so-called gene signature. Such a gene signature is usually a subset of genes contributing to the predictive power (Kim, 2009). Due to abundance of transcripts in a tissue, genes are differently expressed and a tremendous amount of mRNAs can be regarded as noise. Out of more than thousands of genes, the aim of an analysis is, therefore, to select a small subset of candidates such that experimental validation can be performed effectively in order to identify prognostic signatures such as 70- (Van't Veer et al., 2002; Van De Vijver et al., 2002; Buyse et al., 2006) and 76-gene signatures (Wang et al., 2005; Foekens et al., 2006; Desmedt et al., 2007), which are two well-known gene-expression signatures in early stage breast cancer. Most recently, another custom mini-array of 74 genes has been used for gastric cancer prognosis extracted form a microarray signature (Yin et al., 2013), exemplifying the clinical application of dimension reduction by a deterministic approach.

With further advances in HTS technologies, the number of oligonucleotide probes that can be interrogated simultaneously has risen to many thousands and the complexity of analysing the classification performance of the exponentially growing number of gene signatures requires the development of new computational approaches. So far, a number of methods from areas such as data mining, machine learning and pattern recognition have been employed in the analysis of gene expression data derived from high-throughput platforms (Saeys et al., 2007). Here, objects are represented by a vector of features. New techniques have been developed for selecting feature subsets that not only reduce the dimension of the vector without changing the descriptive power for each object, but also to find the minimal subset that maximises the classification performance. In terms of knowledge discovery,

* Corresponding author. Tel.: +44 07474992859.
E-mail addresses: hung-ming.lai@kcl.ac.uk (H.-M. Lai), celal@erciyes.edu.tr (C. Özturk), a.albrecht@mdx.ac.uk (A. Albrecht), kathleen.steinhofel@kcl.ac.uk (K. Steinhöfel).

http://dx.doi.org/10.1016/j.compbiolchem.2015.02.004
1476-9271/

this is based on the parsimony principle (Bell and Wang, 2000), *i.e.*, seeking a model that has as few as possible variables and fits the data sufficiently. Unfortunately, a typical microarray-based cancer experiment consists only of tens to hundreds clinical samples, but each sample has thousands of genes to be considered (Ein-Dor et al., 2006). Therefore, the challenge is to find a model within a tremendous dimensionality by using only relatively few data points. Another challenge is experimental noise. The noise resulting from the experimental design of microarrays, such as the stages of sample preparation, and the hybridisation processes is unavoidable (Tu et al., 2002). Consequently, a number of statistical models have been designed that take into account the quantitative characterization of noise at probe level in order to generate more accurate expression measurements (Irizarry et al., 2003; Wu et al., 2004; McCall et al., 2010). Based on these models, a software tool has been designed for finding single feature polymorphisms at DNA or RNA level by using cross species arrays (Lai et al., 2014).

The general problem of minimal feature subset selection is known to be NP-complete (Davies and Russell, 1994), and a number of heuristics have been proposed for this particular problem. In general, there are three types of feature selection methods for high dimensional gene selection. The three types are filter, wrapper, and embedded techniques. Filter-based techniques measure the discriminative power of genes only according to the intrinsic expression data and statistical evaluation models. Wrapper-based techniques measure the predictive power of gene sets by using machine learning methods and heuristic search strategies. If searching for features is embedded within the update of classification functions or penalty functions during the process of model training, the technique is called an embedded-based method. For an overview on these methods, we refer the reader to (Saeys et al., 2007; Lazar et al., 2012).

Depending on the level of mutually related genes and the degree of mis-representation of physiological and genetic changes caused by the sample set, many different prognostic signatures may exist with the same or varying cardinality. While having a similar classification performance, these distinct signatures may also produce a considerable overlap in patients that are, for example, at high risk of breast cancer recurrence (Fan et al., 2006). This demonstrates that uniqueness of signatures may not be the major concern in the process of identifying gene-expression signatures. Consequently we follow the idea that when calculating prognostic gene signatures it might be advantageous to focus on near-optimal solutions of comparable classification performance, which is in contrast to conventional gene selection methods that aim at the minimal subset of genes having the maximal predictive power (or the minimal misclassification rate).

In this paper, we would like to draw attention to the importance of gene signature evaluation as a subsequent step of gene selection that can be performed by any statistical models or dimension reduction techniques. Along this line, we propose a new approach for establishing gene signature profiles with regard to the number of selected genes, the number of gene-expression signatures and their performance with respect to the generalisation error rate. The new approach not only allows us to compare different gene signatures, but it also provides knowledge about the number and components of signatures having the same cardinality. When analysing biological networks, lists of candidate genes are usually selected according to exploratory data analysis and knowledge about the domain of interest, which is mainly based upon various statistical methods (like *t*-test, ANOVA, SAM (Tusher et al., 2001), LIMMA (Smyth, 2004), PCA, *etc.*). Subsequently, the selected candidate genes are experimentally validated and a gene signature related to a certain disease is identified.

In our approach, we use *t*-statistics in order to simulate the acquisition of candidate genes. Given a set of candidate gene expression measurements, we employ local search methods based upon swarm intelligence, namely, particle swarm optimization (PSO) (Kennedy and Eberhart, 1995, 1997; Kennedy et al., 2001) and the artificial bee colony (ABC) concept (Karaboga, 2005; Karaboga and Ozturk, 2009, 2011; Karaboga et al., 2014) in order to demonstrate the new approach for multiple gene signature evaluation over unrestricted signature cardinalities. For both PSO and ABC, the classification algorithms are designed as randomised swarm wrappers, while the sequential forward selection (SFS) wrapper is deterministic for the baseline construction. We employ the microarray data recording from (Pomeroy et al., 2002) for embryonic tumours in the central nervous system (CNS) in order to demonstrate our approach. The dataset consists of 60 samples with 7129 reference genes. Among the samples 21 are survivors (patients who are alive after treatment) and 39 are failures (patients who succumbed to their disease). As part of our approach, we offer the bead-chain plot for the visual analysis of gene signature profiles, which covers signature sizes, the number of signatures, and the signature performance explored by swarm wrappers.

2. Methods

Our approach aims at the exploration of gene signature evaluations by taking into consideration the signature cardinality, the performance of signatures, and the number signatures. In order to achieve the objective of multiple gene-expression signature analysis, where the particular signatures exhibit a comparable performance, we utilise stochastic optimization methods. Local search procedures based upon swarm intelligence are the method of choice for the evaluation of feature subsets from the search space of all features. By combining randomised search with a classifier, we are able to build a swarm wrapper for finding the most likely solution for a signature at each particular search stage, where multiple search agents are working in parallel. Hiring multiple search agents is at the core of swarm intelligence, which is one of the typical population-based optimisation methods. There are many subtypes of meta-heuristic algorithms being part of the concept of swarm intelligence. In our empirical study, one deterministic wrapper and two meta-heuristics are employed: deterministic sequential forward selection (SFS), particle swarm optimization (PSO), and the artificial bee colony (ABC) concept. While SFS is used for demonstrating that uniqueness of signatures has its limitations, swarm wrappers are designed in order to show the power of multiple putative signatures of comparable performance.

As we intend to develop a method to effectively evaluate signature profiles without restrictions on the signature cardinality, it is important to keep the same size of the population of putative solutions in order to highlight the effect signature cardinalities. If, however, the signature size is constraint, we refer to the problem setting as the *constraint swarm approach*. On the other hand, swarm intelligence methods can innately capture the most feasible solutions without the constraint of having the signature size fixed, which is – we regard this as referred to as the *unconstraint swarm approach*.

2.1. Sequential forward selection (SFS)

Sequential forward selection (SFS) is a stepwise forward procedure. Given a list of candidate genes, a SFS wrapper selects the first gene whose classification performance (evaluated by a given fitness function *F*) is the best. With the inclusion of the first selected gene, *F* evaluates the combined performance of all subsets of two genes and the second gene is selected where their evaluation performance is the best. Iteratively, the third gene is selected with the inclusion of the first two selected genes. The important manner

Please cite this article in press as: Lai, H.-M., et al., A new vision of evaluating gene expression signatures. Comput. Biol. Chem. (2015), http://dx.doi.org/10.1016/j.compbiolchem.2015.02.004