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

# Tumor stratification by a novel graph-regularized bi-clique finding algorithm

Amin Ahmadi Adl[a,*], Xiaoning Qian[a,b,c]

[a] Department of Computer Science and Engineering, University of South Florida, Tampa, FL 33613, USA
[b] Department of Electrical & Computer Engineering, Texas A&M University, College Station, TX 77843, USA
[c] Department of Pediatrics, University of South Florida, Tampa, FL 33620, USA

ARTICLE INFO

*Article history:*
Received 30 December 2014
Accepted 3 February 2015
Available online xxx

*Keywords:*
Tumor stratification
Bi-clique finding
Bi-partite graph
Graph regularization

ABSTRACT

Due to involved disease mechanisms, many complex diseases such as cancer, demonstrate significant heterogeneity with varying behaviors, including different survival time, treatment responses, and recurrence rates. The aim of tumor stratification is to identify disease subtypes, which is an important first step towards precision medicine. Recent advances in profiling a large number of molecular variables such as in The Cancer Genome Atlas (TCGA), have enabled researchers to implement computational methods, including traditional clustering and bi-clustering algorithms, to systematically analyze high-throughput molecular measurements to identify tumor subtypes as well as their corresponding associated biomarkers.

In this study we discuss critical issues and challenges in existing computational approaches for tumor stratification. We show that the problem can be formulated as finding densely connected sub-graphs (bi-cliques) in a bipartite graph representation of genomic data. We propose a novel algorithm that takes advantage of prior biology knowledge through a gene–gene interaction network to find such sub-graphs, which helps simultaneously identify both tumor subtypes and their corresponding genetic markers. Our experimental results show that our proposed method outperforms current state-of-the-art methods for tumor stratification.



## 1. Introduction

One of the important problems in cancer informatics is tumor stratification where the goal is to find the tumor subtypes in a heterogeneous population of samples. Taking advantage of "big" data provided by high-throughput omics profiling techniques, researchers have used standard unsupervised clustering methods such as hierarchical clustering and Non-negative matrix factorization (NMF) to cluster the samples into subtypes (The Cancer Genome Atlas Research Network, 2011; Konstantinopoulos et al., 2008, 2010; Reis-Filho and Pusztai, 2011; Esteva et al., 2005; Hofree et al., 2013). It has been shown that computational methods can help identify tumor subtypes with significantly different survival rates, tumor stages or grades, histological types, and treatment responses. The current efforts mostly have focused on using data from both mRNA expression profiles (The Cancer Genome Atlas Research Network, 2011; Konstantinopoulos et al., 2008, 2010; Reis-Filho and Pusztai, 2011; Esteva et al., 2005) and other somatic mutation profiles for this purpose (Hofree et al., 2013). In order to obtain reproducible and biologically meaningful tumor stratification, there are major challenges to consider for tumor stratification: (1) high-throughput datasets typically contain a relatively small number of samples compared to the number of measurements; (2) high-throughput profiling measurements are often noisy and dynamically vary depending on intra- and extracellular environments; and (3) within the extreme dimension of measured variables in these datasets, it is often conjectured that only a few of them are critical factors or biomarkers and hence traditional clustering based on all the variables may not provide useful biology interpretation.

Recent efforts have been made to address these challenges to improve tumor stratification performance. Many methods based on bi-clustering (Madeira and Oliveira, 2004; Wang et al., 2013; Aradhya et al., 2010) have been proposed for simultaneous identification of both tumor subtypes and their corresponding genetic markers by analyzing profiling data matrices to simultaneously cluster both matrix rows (tumor samples) and columns (genes, for example). The resulting sub-matrices (hence called bi-clusters) correspond to the subsets of tumor samples that behave similarly according to the subsets of genes. For example, a bi-clustering

* Corresponding author. Tel.: +1 8133688686.
*E-mail address:* amin1@mail.usf.edu (A. Ahmadi Adl).

http://dx.doi.org/10.1016/j.compbiolchem.2015.02.010
1476-9271/

algorithm was implemented for stratification of breast cancer tumors using gene expression data in Wang et al. (2013), which shows promising results in identifying phenotypically heterogeneous subtypes together with their associated genes, which can be considered as "personalized" biomarkers for each subtype for more effective disease prognosis, diagnosis, and treatment. Different profile patterns within candidate bi-clusters have been considered and extensively studied in the literature (Madeira and Oliveira, 2004). However, these methods have achieved limited clinical successes, especially when the collected data suffer from the problem of small sample size and high noise (Hofree et al., 2013).

One potential solution strategy for obtaining reproducible knowledge from big data with a small number of uncertain samples is to incorporate prior biology knowledge. One recent breakthrough study by Hofree et al. (2013) has proposed a Network-Based Stratification (NBS) framework. As cellular functions are commonly believed to arise from highly coordinated mechanisms among molecules (Kreeger and Lauffenburger, 2010; Hanahan and Weinberg, 2011), the authors suggested that incorporating prior knowledge provided in gene interaction networks can help improve tumor stratification performance by alleviating the potential limitations of existing data-driven methods, including high noise and small sample size problems in high-throughput data. Furthermore, the authors found that somatic mutation data are less noisy and may be more promising for tumor stratification compared to traditionally used gene expression profiles. They proposed to first smooth somatic mutation data based on a gene-gene interaction network by a graph diffusion method provided in Vanunu et al. (2010). The resulting new matrix is then used for clustering for tumor stratification by Network-Regularized NMF that also incorporates the gene interaction network information (Cai et al., 2008). As in traditional tumor stratification methods, in order to identify important biomarkers related to each subtype, they performed Significance Analysis of Microarrays (SAM) (Tusher et al., 2001) on the network smoothed matrix of mutation profiles after clustering to find genes whose values are significantly different between a subtype and other subtypes. The NBS approach has shown promising results both in identifying subtypes that have significantly different survival rates in ovarian and lung cancers as well as subtypes that are consistent with actual histological types of uterine cancer. They have also shown that the performance is significantly better than other methods such as standard hierarchical clustering and traditional NMF due to the incorporation of gene–gene interaction information. However, tumor stratification and biomarker identification have been done in a separate and sequential manner in NBS and the method is still using a large number of genes for stratification and is unable to directly find related biomarkers for each subtype. Performing clustering and biomarker identification separately may affect the performance. First, the resulting stratification may not be robust when utilizing all the genes for clustering instead of relevant genes, which is biologically more meaningful. Second, biomarker identification is dependent on the quality of the clustering step for stratification, which may not necessarily be accurate and meaningful in the first place.

Motivated by NBS, we develop a novel graph-regularized bi-clique finding algorithm for tumor stratification in this paper. Our new tumor stratification algorithm is specifically designed to address the above mentioned challenges and limitations of existing methods by: (1) incorporating the prior biology knowledge on gene–gene interaction relationships through a graph regularization term to alleviate infamous small sample size and noise problems; and (2) simultaneously clustering tumor samples into subtype groups and identifying genetic markers associated to specific subtypes respectively. We expect that identifying bi-cliques constrained by prior biology knowledge may yield more accurate tumor subtypes as well as robust and personalized biomarkers for corresponding subtypes. Our preliminary experimental results using the Uterine and Ovarian cancer data from TCGA (The Cancer Genome Atlas Research Network, 2011; Hofree et al., 2013) demonstrate that our method is superior or comparable to the performance of the NBS approach, which we consider as the state-of-the-art method for tumor stratification.

## 2. Methods

As in existing tumor stratification methods (The Cancer Genome Atlas Research Network, 2011; Konstantinopoulos et al., 2008, 2010; Reis-Filho and Pusztai, 2011; Esteva et al., 2005; Hofree et al., 2013; Madeira and Oliveira, 2004; Wang et al., 2013; Aradhya et al., 2010), we have a high-throughput profile data matrix $A$, which is an $n \times p$-dimensional matrix containing profiles for $p$ molecules in $n$ available samples. Let $A_{i,j}$ denote the profiled value for the $j$th molecule in the $i$th tumor sample. In this paper, we focus on the somatic mutation data as in Hofree et al. (2013) for fair comparison in our experiments. Hence, $A$ is a binary matrix, in which $A_{i,j} = 1$ indicating that the $j$th gene in $i$th tumor is significantly mutated compared to germ line cells (The Cancer Genome Atlas Research Network, 2011; Hofree et al., 2013). As in NBS, $A$ can be further smoothed by using gene-gene interaction information to obtain a matrix of non-negative real values. With either this binary or smoothed somatic mutation matrix $A$, we are only interested in identifying bi-clusters (or sub-matrices) in $A$ that contains mostly 1s or large values, but not sub-matrices containing all zeros (or small values), which is not indicating a biologically meaningful tumor subtype. This is different from existing bi-clustering methods, often used to analyze microarray expression profiles, which look for bi-clusters with similar patterns no matter whether they have consistently larger or consistently small values. Another major difference in our proposed bi-clique finding algorithm from existing bi-clustering methods is that we only look for a small number of genes that exhibit consistent patterns for each subtype, hoping for biologically more useful and robust biomarkers for consequent disease prognosis and treatment.

In this paper, we will take a graph representation of the given somatic mutation data matrix $A$ so that simultaneous tumor stratification and biomarker identification is transformed to a graph-regularized bi-clique finding problem of searching for densely connected sub-graphs in a bipartite graph constrained with prior biology knowledge on gene–gene interaction information (Fig. 1A). We note that we may extend the method to analyze more general omics profiles including mRNA expression data by first deriving significance tests to obtain a binary matrix capturing the molecule-sample association relationships (The Cancer Genome Atlas Research Network, 2011; Hofree et al., 2013), which requires further modifying the analysis pipeline and thorough performance evaluation and we plan to do it in our future research. Given a somatic mutation data matrix $A$, we first construct a corresponding bipartite graph denoted by $G = (U, V, E)$, where the node set $U = \{u_1, \ldots, u_n\}$ is in one part of the bipartite graph corresponding to available tumor samples; and the node set $V = \{v_1, \ldots, v_m\}$ is in the other part of the graph representing profiled genes. Node $v_i$ is connected to $u_j$ with a weight $A_{i,j}$ (1 or graph-diffused value) if $A_{i,j} > 0$. To achieve tumor stratification, we are looking for a partition of tumor samples with each subset having a corresponding small number of genes that are consistently mutated or have strong association. This can be achieved by finding densely connected sub-bipartite-graphs (relaxed bi-cliques) in the given bipartite graph. There are several existing methods for finding such sub-bipartite-graphs in a bipartite graph (Ding et al., 2006; Barber, 2007; Zha et al., 2001; Dhillon, 2001). Here we propose a novel approach by extending the bi-clique finding method proposed in

Please cite this article in press as: Ahmadi Adl, A., Qian, X., Tumor stratification by a novel graph-regularized bi-clique finding algorithm. Comput. Biol. Chem. (2015), http://dx.doi.org/10.1016/j.compbiolchem.2015.02.010