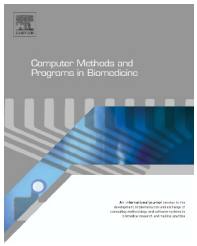
COMPUTER METHODS AND PROGRAMS IN BIOMEDICINE 128 (2016) 69–74

journal homepage: www.intl.elsevierhealth.com/journals/cmpb



InterSIM: Simulation tool for multiple integrative 'omic datasets'

Prabhakar Chalise[*], *Rama Raghavan*[1], *Brooke L. Fridley*[2]

Department of Biostatistics, University of Kansas Medical Center, 3901 Rainbow Blvd, Kansas City, KS 66160, United States

ARTICLE INFO

Article history:
Received 11 September 2015
Received in revised form
23 January 2016
Accepted 18 February 2016

Keywords:
Integrative
Simulation
Clustering
NMF

ABSTRACT

Background and objective: Integrative approaches for the study of biological systems have gained popularity in the realm of statistical genomics. For example, The Cancer Genome Atlas (TCGA) has applied integrative clustering methodologies to various cancer types to determine molecular subtypes within a given cancer histology. In order to adequately compare integrative or "systems-biology"-type methods, realistic and related datasets are needed to assess the methods. This involves simulating multiple types of 'omic data with realistic correlation between features of the same type (e.g., gene expression for genes in a pathway) and across data types (e.g., "gene silencing" involving DNA methylation and gene expression).

Methods: We present the software application tool *InterSIM* for simulating multiple interrelated data types with realistic intra- and inter-relationships based on the DNA methylation, mRNA gene expression, and protein expression from the TCGA ovarian cancer study.

Results: The resulting simulated datasets can be used to assess and compare the operating characteristics of newly developed integrative bioinformatics methods to existing methods. Application of *InterSIM* is presented with an example of heatmaps of the simulated datasets.

Conclusions: InterSIM allows researchers to evaluate and test new integrative methods with realistically simulated interrelated genomic datasets. The software tool *InterSIM* is implemented in R and is freely available from CRAN.

1. Introduction

Identification of molecular subtypes of cancer using high throughput molecular data has been frequently accomplished through the use of clustering [1,2]. Clustering involves the grouping of objects across a disjoint set of classes such that objects within the same class are more similar to one another as compared to the objects in different classes. A large number of clustering methods are available that use a single data type; such as, hierarchical, *k*-means [3], and non-negative matrix factorization (NMF) [4]. In addition to these methods, a few integrative clustering methods have been proposed that utilizes information from multiple data types collected on the same set of samples including: *iCluster* [5], integrative NMF [6], and mixture model based integrative clustering [7].

* *Corresponding author*. Tel.: +1 913 945 7987.
E-mail addresses: pchalise@kumc.edu (P. Chalise), rraghavan@kumc.edu (R. Raghavan), bfridley@kumc.edu (B.L. Fridley).
[1] Tel.: +1 913 945 9412.
[2] Tel.: +1 913 945 5039.

http://dx.doi.org/10.1016/j.cmpb.2016.02.011
0169-2607/

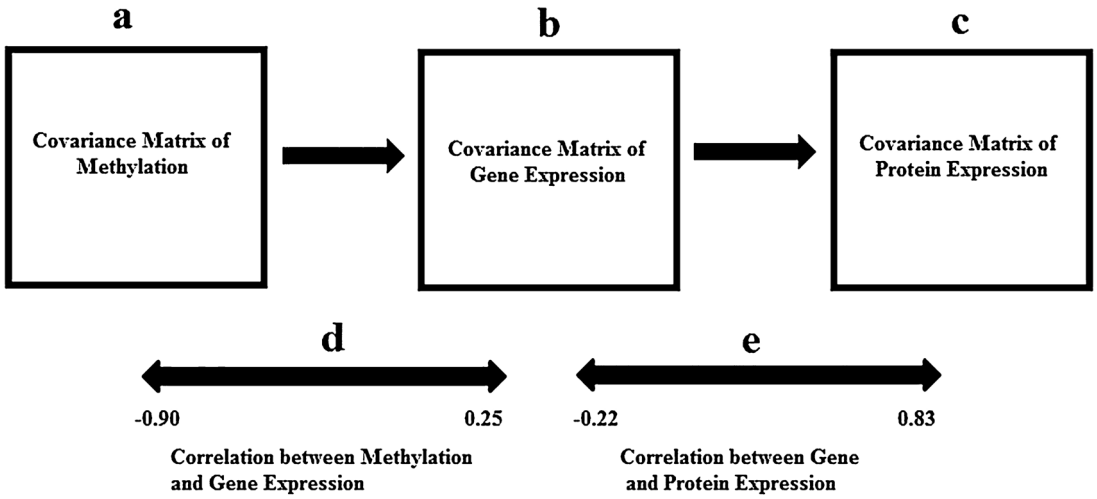


Fig. 1 – Diagram showing the intra- and inter-correlation structure among the features used in the simulation within and between (a) methylation, (b) gene expression and (c) protein expression data from the TCGA studies on ovarian cancer; (d) represents the correlation between the gene level summary of methylation profile and corresponding gene expression (102 pairs were negatively correlated with minimum value of −0.91 and 29 pairs were positively correlated with maximum value of 0.25); (e) represents correlation between the protein expression and corresponding mapped gene expression (14 pairs were negatively correlated with minimum value of −0.22 and 146 pairs were positively correlated with maximum value of 0.83).

A summary of the above mentioned methods can be found in Chalise et al. [8]. However, in order to adequately assess such integrative methods, realistic and interrelated datasets are needed. *InterSIM* bridges this gap by simulating complex interrelated realistic genomic datasets.

Although clustering methods can be used to classify either genes or subjects, the proposed simulation tool focuses on clustering of subjects with the goal of identifying molecular subtypes of disease. In developing the simulation tool, we focused on generating three data types, DNA methylation, mRNA gene expression, and protein expression, on a set of samples with realistic correlation between and within data types. Here are a few examples of the types of relationships we included in the simulation of the data: CpG sites within the same CpG-island would have strong positive correlation, high methylation for a CpG-island upstream of a gene would result in lower mRNA expression or "gene silencing", and higher mRNA gene expression is likely to result in higher downstream protein expression [9]. Such intra- and inter-feature relationships among the data types were based on real data collected on ovarian cancer tumors from The Cancer Genome Atlas (TCGA).

2. Methods

The simulation tool is based on three real datasets from the ovarian cancer study from TCGA – DNA methylation, mRNA gene expression, and protein expression data. In estimation of the relationship in this study, we restricted the tumors to 384 that were common across the three datasets. The level 3 methylation data consists of 27,578 CpG probes from 555 subjects measured using the Illumina 27K, level 3 mRNA gene expression data consists of 17,814 genes from 544 subjects measured with the Agilent G4502A platform, and level 3 RPPA protein expression data contains 187 probes from 412 subjects. Both the methylation and mRNA data were downloaded from https://tcga-data.nci.nih.gov/tcga/tcgaHome2.jsp, and the protein data were downloaded from http://app1.bioinformatics.mdanderson.org/tcpa/design/basic/download.html. Using the CpG to gene annotation and protein to gene annotation information, 367 CpGs and 160 protein probes were found to map to 131 common genes. Based on these three data types measured on 384 subjects with the common mapped features, we estimated the intra- and inter-relationship between the features for use in the simulation of the realistic datasets, Fig. 1.

In simulating the data, we first consider the case where there are no clusters (i.e., only one cluster, $k = 1$, with effect = 0); and then the case where the number of clusters could vary from $k = 2$ to K, where K is the user specified number of clusters. The number of clusters was determined by a handful of features in the various data types. That is, a set of features was selected to have a mean shift in their values so that they would be able to be distinguished among various subgroups or clusters. We start by simulating DNA methylation data, followed by mRNA gene expression, and finally protein expression. Details on the simulation of the data types and the correlation structures are outlined in the following sections.

2.1. *Methylation data*

The methylation β-values at a CpG site, j is the proportion of methylation ranging from 0.0 to 1.0 which are assumed to follow a beta distribution. The logit transformation of such β-values, denoted as M-values, then range from $-\infty$ to ∞ and can be assumed to follow a Gaussian distribution. The