Integration of Multiple Types of DNA Sequence Variations for Gene Activity Impacted

Yang Zhang

Research Institute of Oncology and Hematology (RIOH)
CancerCare Manitoaba
University of Manitoba

Abstract

Objectives: The high throughput next-generation sequencing technology becomes prevalent and the cost dramatically drops, which allows us for sequencing the whole genome (WGS) or whole exome (WES) in cancer studies. From the same WGS or WES data set, four data types of variations are presented: sequence nucleotide variation (SNV) or single nucleotide polymorphism (SNP), structure variation (SV), insert or deletion (INDEL), and copy number variation (CNV). Mining these variations can help to discover cancer drivers and to develop drug to target these genes or mutation. It is difficult to assess where the most affected genes or regions are among all variation events. In this study, we develop a measurement model for integration of the four types of data information. We hypothesize that the four types of variation events work together and ultimately affect the function or activity in the chromosome regions. Specifically, 1) The INDEL mostly destructs the gene DNA sequences including the coding region or regulatory regions; 2) SV will produce abnormal gene products; 3) nonsense or non-synonymous SNV will deactivate the gene activity; 4) More or less than 2 suggest an increase or decrease in gene. All these four variables will produce a gene activity impact score (GAIS). The goal of this study is to develop such an integrative framework of various variation data types for cancer drivers and tumor subtype discovery.
Methods and Results: We present a pairwise hidden Markov model and the probabilistic inference methods to integrate different types of data. In our model, a gene is statistically described as a set of interconnected variables to parameterize model by running Expectation Maximization algorithm. Best parameter set will be obtained to describe the probabilities of the connections between experimental data and state values. A state value is given to the variable following the principle manner of mutation degree of the four data types. Our model is in result of predicting the degrees of impact on the gene activities. The resulting GAIS will be used for clustering analysis. We implemented our model using Matlab. We first tested our model by simulation data. We are now testing breast cancer data. We are analyzing INDEL, SV, SNV, and CNV from WXS data requested from TCGA and estimating the GAIS for around 10 breast cancer patients. We will validate the activity impact scores using matched RNA-seq and normal tissue samples. We hypothesize that the GAIS will reflect in the RNA-seq gene expression level.
Conclusion: In contrast to previous methods that integrate correlated data types such as mRNA and CNV, this is a first study that integrates the uncorrelated data types. The GAIS will help for cancer drivers and tumor subtypes discovery.

Introduction

Modern high-throughput genomics technologies have led to a rapid increase in both the quantity and the variety of functional genomics data. Genomic profiling of multiple data types in the same set of tumors has gained prominence. In a breast cancer study the DNA copy number has been found to relate to gene expression, and 62% of highly amplified genes demonstrated moderately or highly elevated gene expression. The NCI/NHGRI-sponsored Cancer Genome Atlas (TCGA) pilot project is a coordinated effort to explore the entire spectrum of genomic alternations in human cancer to obtain an integrated view of such interplays. For example, an interim analysis of DNA sequencing, copy number, gene expression and DNA methylation data has been conducted in a large set of glioblastomas (TCGA, 2008).

As the high throughput next-generation sequencing technology becomes prevalent and the cost dramatically drops, sequencing the whole genome (WGS) or whole exome (WES) appear in more and more cancer studies. From the same WGS or WES data set, four data types of variations are presented: sequence nucleotide variation (SNV) or single nucleotide polymorphism (SNP), structure variation (SV), insert or deletion (INDEL), and copy number variation (CNV). Mining these variations can help to discover what variation could be the cause of cancer or cancer driver and help to develop drug to target these genes or mutation; It will also help to identify markers that predict the patient's prognosis or treatment benefit. Based on these variation profiling or patterns, we can also cluster tumors into different subtypes.

However, there are millions of these variations in tumor genome compared with normal genome samples. The challenge is how to pinpoint the locations that are affected by sequence variations among the whole genome and how to assess the most relevant variation events among the four types.

One solution is to cluster each data type separately and to find the hot spots in heatmap. However, the hotspots of the four data types mostly are different from each other. It is difficult to assess where the most affected genes or regions are among all variation events. Various integrative clustering approaches have been used for multiple data types and generate a single integrated cluster assignment through simultaneously capturing patterns of genomic alterations. However, all these previously reported methods are for different data types that are related. For example, Qin (2008) performed a hierarchical clustering of the correlation Lee et al. (2008) applied a biclustering algorithm on the correlation matrix to integrate DNA copy number and gene expression data. In both the cases, the goal was to identify correlated patterns of change given the two data types.

While identifying correlated patterns is sufficient for studying the regulatory mechanism of gene expression via copy number changes or epi-genomic modifications, it is not suitable for integrative tumor subtype analysis where the different variations may not be related. For example, the four data types, INDEL, SV, SNV, and CNV from the same WGS or WES sequence data are not related to each other.

In this study, we develop a measurement model for genomic dataset involving more than one data type measured in the same set of tumors. Specifically, we integrate the four types of data information, INDEL, Structure variation (SV), SNP, and copy number variation (CNV). Therefore, the goal of this study is to develop such an integrative framework for cancer drivers and tumor subtype discovery.

We hypothesize that the four types of variation events work together and ultimately affect the function or activity in the chromosome regions, though they are not related to each other. Our hypothesis is based on: 1), The INDEL mostly destructs the gene DNA sequences including the coding region or regulatory regions. One gene or defined bin region could zero to multiple INDELs; 2), SV will produce abnormal gene products; 3), nonsense or non-synonymous SNV will deactivate the gene activity; 4), CNV is 2 for normal diploid chromosomes. More or less than 2 suggest an increase or decrease in gene activity based on previous studies that CNV is correlated to gene expression mRNA levels. All these four variables will produce a gene activity impact score (GAIS). We develop a model for this GAIS estimation.

We propose a novel integrative clustering method called GAIS clustering that is based on a latent variable model. The main idea behind GAIS is that tumor subtypes can be modeled as unobserved (latent) variables that can be simultaneously estimated from INDEL, SV, SNV, and CNV data types. We develop a sparse solution of the GAIS model through optimizing a penalized complete-data log-likelihood using the Expectation Maximization (EM) algorithm (Dempster et al., 1977). The article is organized as follows. In Section 2.1, we xxxx. In Section 2.2, we formulate the xxxx. Then in Section 2.3, we extend the latent variable model to allow multiple data types for the purpose of integrative clustering. Asparse solution is derived in Section 2.4.We demonstrate the method using simulation and lung cancer datasets from published studies in Section 3.

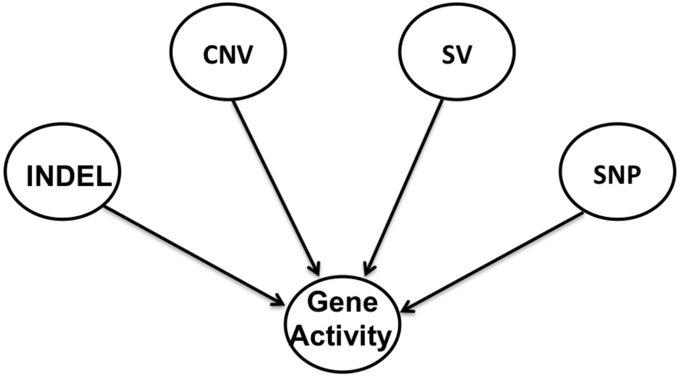
Methods

1 Model

We develop a model named GAIS (Gene Activity Impacted Score) to infer the activities of genes following the patients' data. In this model, we use pairwise hidden Markov model (a subclass of directed graphical model) to estimate the activities of genes for each patient by integrating genetic data. Model GAIS outputs a score GA for each given gene which represents the activity of specific gene comparing with other genes.

In order to describe the biological interactions between target molecules in a signal gene by pairwise Markov random field, we first include all target molecules and interactions into directed graph. In the directed graph of a protein-coding gene, each node represents the target molecule such as the copy number vari-

Figure 1: Directed graph of a protein-coding gene.



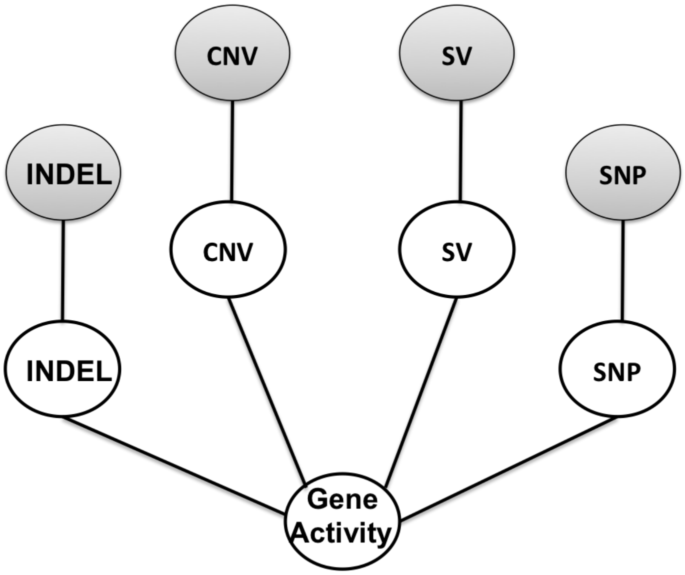
The arrows represents the interactions between nodes. The parents nodes which represent target molecules: INDEL, CNV SV, SNP affecting the offspring node Gene Activity unilaterally and collaboratively.

ation, structural variation etc. . Each arrow represents the dependency between these target molecules. Figure 1 demonstrates the directed graph of a protein-coding gene with its target molecules. In the diagram, we use nodes (circles) to represent target molecules: INDEL, CNV, SV, SNP (representing insert deletion, copy number variation, structural variation and single nucleotide polymorphism respectively) and gene activity. All arrows describe the interactions between each molecule, specifically, CNV has a positive influence on Gene Activity, while INDEL, SV and SNP have negative influence on Gene Activity. These influences will be introduced later to calculate the final score of gene activity.

In pairwise hidden Markov model, variables can be described as the pairwise of observations and hidden. Due to the four types of data (INDEL, CNV, SV and SNP), Figure 1 can be presented by pairwise Markov model graph, which is demonstrated in Figure 2. From experiments, we collected four types of data from patients to estimate the observation variables in our hidden Markov model. While we set up five hidden variables at the same time. More details are introduced in Figure 2. Note that the interactions represented by directed graph (Figure 1) that do not change in Figure 2 (e.g. variables INDEL, CNV, SV and SNP are consistently independent and their influences on Gene Activity are kept as same as in Figure 1).

The pairwise hidden Markov model encodes the activity of a given protein-coding gene by picking up random variables. To develop a graphical probabilistic model, we set up hidden variable $X = \{X_1, X_2, X_3, X_4, X_5\}$ to represent molecules INDEL, SV, SNP, CNV and Gene Activity. Observation variable $Y = \{Y_1, Y_2, Y_3, Y_4\}$ represents the data of INDEL, SV, SNP and CNV. Thus, if we observe $X = x$ and $Y = y$, the pairwise hidden Markov model (Figure 2)

Figure 2: Pairwise Markov random field of a protein-coding gene.



The edges represent the interactions between nodes. The grey circles represent the observed data of INDEL, CNV, SV and SNP. The white circles are the five hidden variables of INDEL, CNV, SV, SNP and Gene Activity. All the independencies (no arrows between INDEL, CNV, SV etc.) shown in Figure 1 are present in Figure 2 as well. The parents nodes which represent variables INDEL, SV, SNP still keep negative influences on the offspring node Gene Activity, and the positive influence from CNV on Gene Activity is also kept in Figure 2.

has the complete data likelihood function of a given gene as follows:

$$g(x, y, \theta) = P_\theta\{X = x\}P_\theta\{Y = y|X = x\}, \tag{1}$$

where $\theta = (k, g_1, g_2, g_3, g_4)$ is parameter set of the pairwise hidden Markov model. The definition of each parameter is given:

$$\begin{aligned} &P\{X_i = Y_i|X_i = v_i\} = k, \\ &P\{X_i = X_5|X_5 = v_5\} = g_i, \quad \text{for } \ i \in [1, 4]; \end{aligned}$$

where $v_1 = v_2 = v_3 = v_4 = \{0, 1, 2, 3, 4\}$. We assume that the parameter k is the probability when the observed data and realistic molecule are consistent, and parameter g_j $(j = 1, 2, 3, 4)$ is the probability when the variable such as INDEL, SV, SNP, CNV has a influence on Gene Activity with taking the same state value. Now, Equation 1 can be written:

$$g(x, y, \theta) = \prod_{i=1}^{4} k^{I\{x_i=y_i\}}(1-k)^{I\{x_i\neq y_i\}} g_i^{I\{x_i=x_5\}}(1-g_i)^{I\{x_i\neq x_5\}}, \tag{2}$$

equivalently, the logarithm function of Equation 2 can be rewritten:

$$\begin{aligned} log\ (g(x,y,\theta)) =&\sum_{i=1}^{4} N_i log(k) + (4-N_i)log(1-k) + M_i log(g_i) \\ &+(1-M_i)log(1-g_i), \end{aligned} \tag{3}$$

where $N_i = I\{X_i = Y_i\}$ and $M_i = I\{X_i = X_5\}$ for $i \in \{1,2,3,4\}$.

2 Expectation Maximization algorithm

To learn the parameters of the model, we apply the Expectation Maximization (EM) algorithm (Dempster *et al.*, 1977). EM algorithm is an iterative procedure that obtains the parameters by iterating the probabilities of hidden variables and changing parameters to maximize the expectation of its logarithm likelihood function. In general, EM algorithm has two steps to generate the iterative parameters, the detail of procedure is given as follow:

E-step: Define Q function: $Q(\theta|\theta_{\text{old}}) = E_{\theta_0}(logP(X,Y|\theta))$, an expectation function of its logarithm likelihood function with hidden variables for the given complete data.

M-step: Find the parameter θ in the next iteration, which is defined as $\theta = arg\{\max_\theta Q(\theta|\theta_{\text{old}})\}$, or at least make sure $Q(\theta) > Q(\theta_{\text{old}})$ (A.K.A generalized EM algorithm).

Going back to our model, by applying EM algorithm, we obtain the Q function based on Equation 3 given as follows:

$$\begin{aligned} Q(\theta|\theta_0) =&E_{\theta_0}(log(g(X,\theta))|Y=y) \\ =&\sum_{i=1}^{4} E_{\theta_0}(N|Y=y)log(k) + E_{\theta_0}(4-N|Y=y)log(1-k) \\ &+E_{\theta_0}(M_i|Y=y)log(g_i) + E_{\theta_0}(1-M_i|Y=y)log(1-g_i). \end{aligned}$$

Where Y is the fixed data, parameter θ_0 is the initial condition.

Thus, in order to maximize the above function, we set $\frac{\partial Q(\theta_1|\theta_0)}{\partial\theta} = 0$ to learn the next iteration parameter θ_1. The following equations about parameters in next iteration can be obtained:

$$\begin{aligned} k=& \frac{1}{4}E_{\theta_0}(N|Y=y) \\ g_i=& E_{\theta_0}(M_i|Y=y),\ for\, i \in \{1,2,3,4\} \end{aligned}$$

Repeat the above steps, we will have the best parameters θ for each gene of a patient, which makes the likelihood changes less than 0.01%. In order to connect the entire data distribution, we consider to cluster genes parameters by Hierarchical method into 9 different subgroups (estimation the number of clusters by gap statistic method [2]). In each group, we average the corresponding parameters to have one best parameter set. After 9 sets of best parameter obtained, they are used to calculate the Gene Activity Impacted Score in the next section.

3 Gene Activity

We calculate the Gene Activity Impacted Score (GAIS) for each given gene by considering the influences between target molecules. The definition of GAIS is based on three factors: the best parameters obtained by EM algorithm, the influences from four different type molecules (INDEL, SV, SNP, CNV) on Gene Activity and the state values of observation data. Since we introduced the best parameter sets from 9 subgroups by clustering, $i \in [0, 9]$. We use the linear combination equation to express GAIS, which is given as follows:

$$GAIS = \frac{1}{N_m} \sum_{i=1}^{N_m} \sum_{j=1}^{3} k_i^{*m} g_{i4}^{*m} y_4 - k_i^{*m} g_{ij}^{*m} y_j,$$

where N_m is maximum size of the clustered gene subgroup m, k_i^{*m} is the optimized parameter k obtained by EM algorithm for the gene in subgroup m. The observation data $Y = \{Y_1 = y_1, Y_2 = y_2, Y_3 = y_3, Y_4 = y_4\}$ represents the state values of molecule INDEL, SV, SNP, CNV respectively. The best averaged Parameter sets reflects the probabilities of which observation data become the weighted state values in our model. Note that GAIS indicates the comparison between genes. The score of GA might be very small for a single gene, but the scores for all genes are in the same magnitude. For the convenience of cluster later, we enlarger the score into $[-10^2, 10^2]$.

Results

1 Simulation

2 Lung cancer whole genome sequence data

Discussion

References

[1] **Vaske C J, Benz S C, Sanborn J Z, Earl D, Szeto C, Zhu, J, Haussler D** and **Stuart J M**. *"Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM"*. Bioinformatics. 26 (12): i237- i245 2010.

[2] **Robert T Guenther W and Trevor H** *"Estimating the number of clusters in a data set via the gap statistic"*. J. R. Statist. Soc. 63(2): 411-423 2001.