MADGiC: An R package for identifying driver genes in cancer

Keegan Korthauer and Christina Kendziorski

Contents

1 Introduction

We introduce a statistical method that focuses on data integration in the realm of cancer bioinformatics. Motivated by the enormous amount of genetic data made publicly available by databases such as The Cancer Genome Atlas (TCGA) project and the Catalogue of Somatic Mutations In Cancer (COSMIC), we aim to identify mutated genes that contribute to the disease process (driver genes). The main challenge in identifying driver genes is that cancer genomes rapidly accumulate benign mutations (passengers) after tumor initiation, and thus the passenger mutations greatly outnumber the drivers. Further complicating matters, passenger mutations do not occur at a uniform rate throughout the genome.

Early statistical methods for identifying driver genes in cancer relied primarily on frequency-based criteria (i.e. identifying driver genes as those showing higher mutation rates than expected by chance). However, more recent studies have identified many other properties of drivers such as increased functional impact, enrichment for specific types of mutations, and highly structured spatial patterns. Though tools exist for probing some of these factors one at a time, we have developed a unified framework to identify driver genes that incorporates all of these criteria. This is done by jointly modeling the mutational event with its functional impact, as well as incorporating the mutational enrichment and patterns as prior information. The method is called MADGiC, a Model-based Approach for identifying Driver Genes in Cancer, and shows substantially increased power (with a well controlled false discovery rate) compared to competing methods in simulation studies. Further advantages are demonstrated in case studies using data from the TCGA ovarian and lung cancer cohorts. For more details on this method, see Korthauer and Kendziorski [2015].

2 Citing the software

Please cite the following article when reporting results from the software:
Korthauer, K. D., and Kendziorski, C. MADGiC: a model-based approach for identifying driver genes in cancer. *Bioinformatics*, 2015.

3 The model

3.1 Driver gene model framework

Our primary aim is to prioritize genes that have been somatically mutated in cancer based on the likelihood that they are driver genes. A driver gene is defined as a gene harboring a mutation that provides a selective advantage to the cancer cell. The empirical Bayesian hierarchical mixture model framework we develop considers three sources of evidence for driver activity: (1) increased frequency of mutation compared to a gene-specific background mutation model, (2) evidence of functional impact, and (3) a non-random spatial pattern of mutations.

3.2 Generative model

Consider a single gene indexed by g, from a total of G genes having at least one nonsilent somatic mutation. Note that nonsilent mutations include missense mutations, frameshift indels, and in frame indels. Further consider an independent sample of size J, indexed by j, each with at least one nonsilent somatic mutation in one or more of the G genes. Let $\vec{X}_g = X_{1g}, ..., X_{Jg}$ be the vector of observed nonsilent mutation states of gene g for all samples (where $X_{jg} \in \{0, 1\}$ and $1 =$ one or more

nonsilent mutations anywhere in the gene; 0 = no mutations in the gene). Next, let $\vec{S}_g = S_{1g}, ..., S_{Jg}$ be the vector of functional impact scores for mutations in gene g for all samples. Finally let $Z_g \in \{0, 1\}$ be the indicator that gene g exhibits driver activity.

We are interested in the posterior probability that gene g is a driver gene given the mutation status and impact score for that gene across J independent samples:

$$P(Z_g = 1|\vec{S}_g = \vec{s}, \vec{X}_g = \vec{x}) = \frac{P(Z_g = 1)\prod_{j=1}^{J} P(S_{jg} = s_j, X_{jg} = x_j|Z_g = 1)}{\sum_{k\in\{0,1\}} P(Z_g = k)\prod_{j=1}^{J} P(S_{jg} = s_j, X_{jg} = x_j|Z_g = k)} \tag{1}$$

We assume that the presence of mutations in gene g and sample j depends on driver status. Specifically, $X_{jg}|\, Z_g = z \sim Bern((1-z)b_{jg} + zd_g)$ where $b_{jg} \in (0, 1)$ is the background (passenger) mutation probability for sample j, gene g, and $d_g \in (0, 1)$ is the driver mutation probability for gene g. To enforce that the driver mutation probability is at least as high as the average passenger mutation probability (i.e. that $d_g > \bar{b}_{.g}$), we let $d_g \sim Beta(\alpha, \beta)$ truncated below at $\bar{b}_{.g}$.

Likewise, we assume that the distribution of functional impact scores across all genes and all samples depends on driver status. Specifically, $S_{jg}|\, X_{jg} = 1, Z_g = z \sim (1-z)f^p + zf^d$, where f^p is the distribution of functional impact scores for passenger genes and f^d is the distribution of functional impact scores for driver genes. Note that we are assuming a common functional impact score profile for all driver mutations, and another for all passenger mutations, independent of mutation frequency.

The prior predictive distributions $P(\vec{S}_g = \vec{s}, \vec{X}_g = \vec{x}|Z_g = 1)$ and $P(\vec{S}_g = \vec{s}, \vec{X}_g = \vec{x}|Z_g = 0)$ are obtained by averaging over the prior distribution of d_g. Then,

$$P(\vec{S}_g = \vec{s},\vec{X}_g = \vec{x}|Z_g = 1) = \frac{B(\alpha^*, \beta^*)[1 - F_{(\alpha^*,\beta^*)}(\bar{b}_{.g})]}{B(\alpha, \beta)[1 - F_{(\alpha,\beta)}(\bar{b}_{.g})]} \prod_{j=1}^{J} \hat{f}^d(s_j)^{x_j}$$

$$P(\vec{S}_g = \vec{s},\vec{X}_g = \vec{x}|Z_g = 0) = \prod_{j=1}^{J} (\hat{f}^p(s_j)\hat{b}_{jg})^{x_j}(1 - \hat{b}_{jg})^{1-x_j}$$

where $F_{(\alpha,\beta)}$ is the cumulative distribution function of the beta distribution with shape parameters (α, β); B is the Beta function; $\alpha^* = \sum_{j=1}^{J} x_j + \alpha$; and $\beta^* = J - \sum_{j=1}^{J} x_j + \beta$. Then the final form of the posterior probability is easily obtained from Equation 1.

4 Example data from TCGA

Data from the TCGA ovarian cancer cohort (downloaded from the TCGA data download portal at `https://tcga-data.nci.nih.gov/tcga/` on 10/1/2013) are included

as an example. In this collection of 463 ovarian cancer samples there are 5,849 silent mutations (mutations that do not alter the amino acid sequence) located in 4,369 genes and 21,800 nonsilent mutations (mutations that cause a change in the amino acid sequence) located in 10,164 genes. The median (range) total number of mutations per sample is 60 (1-209). For silent mutations, the median (range) is 11 (0-51), and 41 (0-175) for nonsilent mutations. There is very little positional overlap of mutations across samples and only 62 genes have a nonsilent mutation in more than 10 samples.

5 Quick start

5.1 Installation

Download the .tar.gz package source file. There are two options for installation. One option is through the command line. To do so, cd into the directory where the .tar.gz package source file is saved, and enter the command

```
R CMD INSTALL MADGiC_X.tar.gz
```

where 'X' is the version number that matches the source file name (e.g. 0.1).

Another option is through a GUI (graphical user interface), such as RStudio. To do so using RStudio, go to the 'View' menu and click on 'Show Packages' (or simply click on the 'Packages' tab of the lower right-hand window under the standard view settings). Then click on 'Install Packages'. Under the 'Install from' dropdown menu, choose 'Package Archive File'. Finally, navigate to the file folder where you have saved the .tar.gz package source file. Click 'Install' (make sure the 'Install Dependencies' option is checked).

5.2 Running MADGiC on example data

Once the package is successfully installed, it will need to be loaded into the R session using the library command:

```
> library(MADGiC)
```

To fit the model to the example data set from TCGA ovarian using default settings, first create a pointer to the MAF file to be analyzed:

```
> maf.file <- system.file("data/OV.maf",package="MADGiC")
```

Then use the `get.post.probs` function to estimate posterior probabilities that each gene is a driver:

```
> post.probs <- get.post.probs(maf.file)
```

5.3 Running MADGiC other input data

To run MADGiC on other datasets, it is helpful to know what inputs the `get.post.probs` function will take and which arguments can be customized.

5.3.1 Required Input

MAF data file: `maf.file` is a pointer to an MAF (Mutation Annotation Format) data file containing the somatic mutations. Currently, NCBI builds 36 and 37 are supported.

5.3.2 Optional Input

If alternate sources of expression level or replication timing are present, these can be used by specifying pointers to the following:

Expression data file: `expression.file` is a pointer to a .txt file containing gene expression data if user wishes to supply one (default is to use an average expression signal of the CCLE). The .txt file should have two columns and no header. The first column should contain the Ensembl Gene ID (using Ensembl 54 for hg18) and the second column should contain the expression measurements. These can be raw or log-scaled but should be normalized if normalization is desired.

Replication data file: `replication.file` is a pointer to a .txt file containing replication timing data if user wishes to supply one (default is to use data from Chen et al. (2010)). The .txt file should have two columns and no header. The first column should contain the Ensembl Gene ID (using Ensembl 54 for hg18) and the second column should contain the replication timing measurements.

5.3.3 Customizable arguments

Number of simulated datasets: `N` is the integer number of simulated datasets to be used in the estimation of the null distribution of functional impact scores. The default value is 20

Hyperparameters α and β: `(alpha, beta)` are numeric values of first and second shape parameters, respectively, of the prior Beta distribution on the prob- ability of mutation for driver genes. Default value of (0.2, 6) is chosen as a compro- mise between a cancer type with a relatively low mutation rate (Ovarian cancer, fitted value from COSMIC of (0.15, 6.6)) and one with a comparatively high mutation rate (Squamous cell lung, fitted value from COSMIC of (0.27, 5.83)), but results are robust to changes in this parameter. Note that intuitively (and empirically), a higher

mutation rate overall leads to a higher driver mutation rate overall - and thus less mass is concentrated in the left tail of the distribution.

References

K. D. Korthauer and C. Kendziorski. Madgic: a model-based approach for identifying driver genes in cancer. *Bioinformatics*, page btu858, 2015.