A proposal of a graph-based computational method for ranking significant set of related genes in cancer

Jorge Francisco Cutigi[1,2], Adriane Feijó Evangelista[3], Adenilso da Silva Simão[2]

[1]Instituto Federal de Educação, Ciência e Tecnologia de São Paulo – IFSP São Carlos
São Carlos – SP – Brasil

[2]Instituto de Ciências Matemáticas e de Computação – Universidade de São Paulo
São Carlos – SP – Brasil

[3]Centro de Pesquisa em Oncologia Molecular – Hospital de Câncer de Barretos
Barretos – SP – Brasil

cutigi@ifsp.edu.br, adriane.feijo@gmail.com, adenilso@icmc.usp.br

Abstract. *Identifying significant mutations in cancer is a key point in Cancer Genomics, and it is one of the biggest challenges in the area. Computational methods for identifying significant mutations have been developed in recent years. In this work, we present a proposal of a flexible computational method with an extensive biological base for ranking significant set of related genes in cancer. Our method considers data about mutations, type of mutations, gene interaction networks and mutual exclusivity pattern.*

1. Introduction

Cancer is known to be caused by the accumulation of genetic alterations during the individual's life, from small changes in nucleotides to more considerable variations in genetic material. These changes are called genetic mutations, which have been studied for several period of years, by the DNA/RNA sequencing. Next-Generation Sequencing (NGS), promote fast and cost-effective genomic sequencing. It enables the generation of large volume of cancer data in a short time. Due to the abundance of genomic data, there is a challenge to process NGS data for useful clinical information. In this sense, clinical bioinformatics develops and uses computational methods for the interpretation of data.

One of the categories of computational methods includes those that intend to identify significant mutations (or driver mutations) for cancer. A cancer cell may have two types of mutations: 1) passenger mutations, which are mutations that do not change the cell behavior; and 2) driver mutations, which are mutations that cause harmful behavior in the cell and are responsible for cancer, i.e., they give the cells a selective advantage in comparison to the other cells, increasing their survival and reproduction [Stratton 2009].

The identification of driver mutations is crucial in cancer genomics and it is one of the most significant challenges in the field [Hou and Ma 2013, Raphael et al. 2014, Cheng et al. 2015]. The research of computational methods and their algorithms for the identification of significant mutations in cancer has developed a lot in recent years. Each method presents specific computational and biological characteristics. These methods have found new associated cancer genes, from a single gene to a set of them.

In this context, the objective of this work is to propose a computational method for ranking set of related genes significantly mutated in cancer. We merged computational

and biological perspectives in order to create a method that can achieve significant results for the study of cancer genomics. The main advantage of our method is the use and manipulation of more biological data, and the flexibility in considering this in the analysis. The proposed method presented here is part of a work in progress. The method was defined in a theoretical way with professionals of the Computer Science and Cancer Genomics. The implementation, test, and evaluation of the results are still in development.

2. Related work

Computational methods for identifying significant mutations in cancer have been proposed since the emerging amount of sequencing data. Most of the methods have been proposed since the beginning of the decade. Some methods are related to each other, having some similarities and differences among them. In this section, we briefly describe some methods, showing their main characteristics.

The HotNet method [Vandin et al. 2011] finds multiple sets of genes that are mutated in a relatively high number of patients. The algorithm creates a gene graph and uses the network diffusion algorithm to separate nodes (genes) and find relevant subsets. As a criterion to find these subsets, it is defined how hot is a gene (how is your interaction in the network) and his coverage (how many times it was mutated in the patients). The HotNet2 method [Leiserson et al. 2015] extends the HotNet method. The diffusion process used in this method better encodes the network topology, and the method uses a directed graph to find significant subnetworks. The Hierarchical HotNet method [Reyna et al. 2018] uses gene network and gene scores to construct a hierarchy topologically close and high-scoring subnetworks [Reyna et al. 2018].

The Dendrix method [Vandin et al. 2012] works with the hypothesis that pathways contain a set of genes that the mutation has a high coverage (most patients have at least one mutation in the set) and exhibits a high pattern of high exclusivity (most patients have only one mutation in the set). Dendrix method measures how much a set of genes has these characteristics defining a weight function to get this measure. Based on the measure, Dendrix method uses two approaches to select one set of genes that get the best measure. The Multi-Dendrix method [Leiserson et al. 2013] uses the same weight function than Dendrix method and it is also capable to find multiple driver pathways.

The MEMo method [Ciriello et al. 2012] performs statistical analysis and tests to identify characteristics in gene network modules based on three criteria. The objective of the MEMo method is to identify the set of related genes (modules) that are frequently altered, belonging to the same biological process with pattern of mutual exclusivity.

3. Biological concepts

In genomics, there are complex interactions around genes and their produced proteins, which are called gene interaction networks. In this kind of networks, genes are the nodes, and the edges connect the genes that are physically interacting or functionally related. Gene networks are largely used in Cancer Genomics, and there are many databases with gene interaction information, for example the ReactomeFI [Fabregat et al. 2018].

Mutations in cancer occur in different scales, from a simple variation of a single nucleotide to a huge alteration in a significant part of the chromosome or even in the whole

chromosome. SNVs (Single Nucleotide Variants) happens when a single nucleotide is substituted by the other. SNVs, can be *missense* or *nonsense*. Missense mutation results in the substitution of one amino acid for another. A nonsense mutation creates a signal to the cell stop building the protein, resulting in a shortened protein. InDels (Insert and Deletions) happens when a single or small sequence of nucleotides can be inserted or deleted of part of the DNA sequence. InDels cause *frameshift* mutations, which leads for errors in the reading sequence to produce the protein. *Splice* mutations are important mutations that occur in the regions where the splicing phase happens, in the creation of messenger RNA. Some databases provide important data about known relevant cancer genes and mutations, for example the MSigDB [Subramanian et al. 2005].

The mutual exclusivity pattern is related to a group of two or more genes, where the genes into this group are rarely mutated in the same patient, i.e., simultaneous mutations of these genes in the same patient are less frequent than they are expected by chance [Kim et al. 2017]. On the other hand, different genes of the group can be mutated in different patients.

4. Method

In order to find a group of related genes significantly mutated in a cohort of patients, we propose a network-based method to rank a set of genes with size k. Our method has six steps and uses somatic mutation data, gene interaction network and mutual exclusivity pattern. With these data, we define a function to find and rank a set of related genes that are mutated in most patients with a high score and present a pattern of mutual exclusivity. In Figure 1 we present an overview of the method.

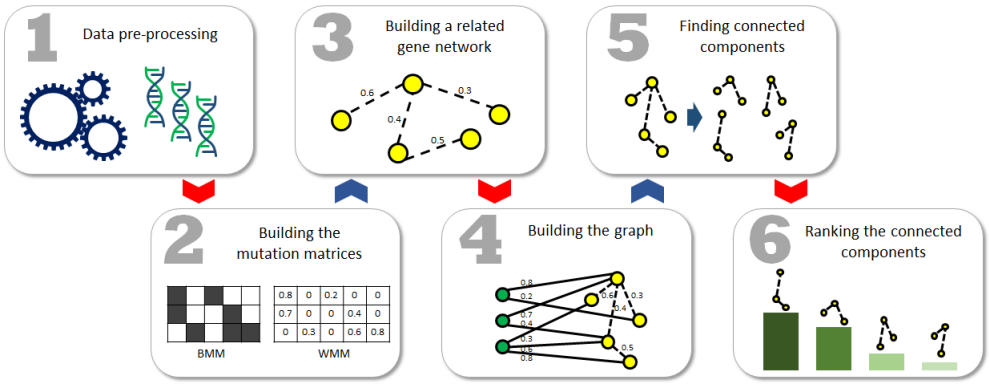

Figure 1. An overview of the proposed method.

Step 1: Data pre-processing In this first step, we pre-process genomic data in order to keep the most relevant genes in the analysis. For this, we keep genes mutated in at least 2% of the patients. Furthermore, we also assign a score for each pair patient-gene, based on its type of mutation, and if the gene is known to be relevant for cancer, extracted from the MSigDB database. For each mutation in the gene, the score is summed by 0.5 if the mutation was nonsense or frameshift. If the mutation is missense, splice or in-frame, it is summed 0.2. After that, the final sum is divided by the number of mutations of the gene. Finally, if the gene is known to be relevant in MSigDB, we sum 0.5 in the score. In summary, all pairs of a mutated gene g_j in a patient p_i has a score $s(p_i, g_j)$ that represents how important that mutation can be in that patient.

Step 2: Building the mutation matrices In this step, we build two matrices: 1) Weighted Mutation Matrix (WMM), and 2) Binary Mutation Matrix (BMM). In these matrices, the rows are patients and columns are genes. In a WMM matrix wm, the entry wm_{ij} has the value $s(p_i, g_j)$ obtained in Step 1, i.e., the wm matrix is a real value matrix. After that, we can derive a BMM matrix bm, where the entry bm_{ij} has value 1 if wm_{ij} is greater than zero, and 0 otherwise.

Step 3: Building a related gene network In this step, we build a Related Gene Network (RGN), that represents genes which are functionally related. Our method uses a gene interaction network from known databases, as ReactomeFI. From the ReactomeFI network, we build the RGN. We define a threshold γ, and for each gene pair $\{g_i, g_j\}$, we derive a measure using the Jaccard coefficient. For this measure, two genes g_i and g_j are considered proximal if they share a large number of common neighbors. Considering the set of neighbors of g_i and g_j as $N(g_i)$ and $N(g_j)$, respectively, we have the Jaccard coefficient calculated as follows: $J(g_i, g_j) = \frac{|N(g_i) \cap N(g_j)|}{|N(g_i) \cup N(g_j)|}$. If the measure is greater than γ, then we placed an edge between g_i and g_j, with the measure as its weight. This process is similar to the approach presented by [Ciriello et al. 2012], where the authors use the Jaccard coefficient as the measure, and $\gamma = 0.02$.

Step 4: Building the graph In this step, we build a graph $\mathcal{G}$, where the right side of the graph are the genes, and the left side are the patients. To build $\mathcal{G}$, we use the wm and bm matrices obtained in Step 2, and the gene network RGN obtained in Step 3. In the bm, when a mutation is observed in a gene g_j in a specific patient p_i, an edge is included in $\mathcal{G}$ linking the patient p_i with the gene g_j in RGN. It is defined a weight in this created edge, where the weight is the corresponding entry wm_{ij}. In Figure 2, we present a small graphical example of the graph.

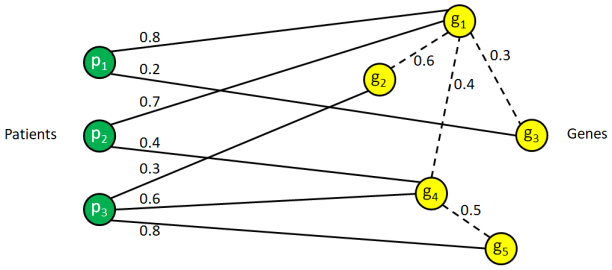

Figure 2. An example of the graph.

Step 5: Finding connected components After the Step 4, we have a graph $\mathcal{G}$ with important information contained in itself. Now, in the Step 5, we find a set cc of connected components of size k in the gene nodes of the graph. A $k-$size connected component in a graph is a subgraph of size k, where there is at minimum one path between all the pair of the k nodes, i.e., all nodes of the subgraph are connected to each other by their paths. For example, in the Figure 2, considering $k = 3$, the component $c_1 = \{g_1, g_2, g_3\}$ is connected, but $c_2 = \{g_1, g_3, g_5\}$ is not.

Step 6: Ranking the connected components First, in Step 6, we rank the components found in the later step. The ranking is based on three aspects, considering the set of the genes in one component c of the set cc. For this, we use three aspect functions $I(c)$, $R(c)$, and $W(c)$. For the next examples, it is considered the graph of the Figure 2 and the component $c = \{g_1, g_2, g_3\}$:

- The relevance of the mutation in the pair gene-patient: for this aspect, we define the function $I(c)$, that is the sum of the edge weights that link genes and patients. For the running example we have that $I(c) = 0.8 + 0.7 + 0.3 + 0.2 = 2.0$.
- How related are the genes: in this aspect, we define the function $R(c)$, that is the sum of the weights of the edges of c, considering only the gene nodes of the graph and their edges. For the running example we have that $R(c) = 0.6 + 0.3 = 0.9$.
- Mutual exclusivity pattern: to evaluate and quantify the mutual exclusivity among the genes in the component, we use the weight function $W(G)$ defined in the Dendrix method [Vandin et al. 2012]. The function considers the trade-off between the coverage and exclusivity of a set of genes, looking for mutual exclusivity in the BMM. The weight function is defined by $W(G) = 2|\Gamma(G)| - \sum_{g \in G} |\Gamma(g)|$, where $\Gamma(g)$ are the patients where g is mutated and $\Gamma(G)$ are the patients where at least one gene in G is mutated. For the running example we have that $W(c) = 2$.

We apply the aspect functions for all components of cc, then we normalized all results in a scale between 0 and 1. Next, we define a rank function $F(c)$ which considers all the aspects presented, i.e., the rank function combines the result of these three normalized aspect functions. For each aspect function, we have to choose a parameter, in order to define the weights of each aspect. In this way, the rank function $F(c)$ is defined as $F(c) = \alpha I(c) + \beta R(c) + \delta W(c)$. With this, we can define the parameters α, β, and δ according to the interest of the investigation. For example, if we are more interested in considering the mutual exclusivity pattern, it is expected to define the δ parameter greater than others. This makes the method flexible. Finally, for each component c in the set of components cc, the rank function $F(c)$ is applied. Then, according to the result of the function, the set of the genes is sorted in decreasing order.

Experimental evaluation To evaluate our method, we will use real and artificial genomic data. The real data is necessary for biological evaluation, i.e., to know if the method outputs results that make sense from a biological perspective. Artificial genomic data may be used to analyze the statistical significance of the results, and assert if the method can find the expected outcomes. We propose the experimental evaluation in three ways: 1) Evaluation via real data: analyze how good is the method to identify known and new significant genes for cancer, using known benchmarks and the expert analysis; 2) Evaluation via statistical analysis: perform the statistical analyses used in the HotNet family methods; and 3) Comparison to other methods: compare to existing computational methods, both from biological and computational perspectives.

5. Conclusion

This paper describes a proposal of a computational method for ranking significant set of related genes in cancer. The method takes advantage of many kinds of data, as the type of mutation for a gene in a patient, the interaction among genes, and the pattern of mutual exclusivity. The method have parameters that can be adjusted according to the goal of the analysis. Our main contribution is to put these data together and consider all of them in the identification of possible significant set of genes in cancer using a graph. In comparison with other methods, our method could better fit in the same category of MEMo and MEMCover, although the method uses ideas from Dendrix and HotNet. We expect that our method could have good results, because it uses some important strategies performed by other classical and relevant methods, making an ensemble of these strategies.

The next steps are the implementation of the method and, eventually, make some modifications in the presented steps. As possible improvements, we can cite add new parameters to create the WMM in Step 1, in order to define the scores, like mutational signatures and information about cancer hotspots. Another improvement is to try new strategies to build the RGN. We could define statistical analysis to better defines the thresholds and the parameters of the rank function.

References

Cheng, F., Zhao, J., and Zhao, Z. (2015). Advances in computational approaches for prioritizing driver mutations and significantly mutated genes in cancer genomes. *Briefings in Bioinformatics*, 17(4):642.

Ciriello, G., Cerami, E., Sander, C., and Schultz, N. (2012). Mutual exclusivity analysis identifies oncogenic network modules. *Genome Res*, 22.

Fabregat, A., Jupe, S., Matthews, L., Sidiropoulos, K., Gillespie, M., Garapati, P., Haw, R., Jassal, B., Korninger, F., May, B., Milacic, M., Roca, C. D., Rothfels, K., Sevilla, C., Shamovsky, V., Shorser, S., Varusai, T., Viteri, G., Weiser, J., Wu, G., Stein, L., Hermjakob, H., and D'Eustachio, P. (2018). The reactome pathway knowledgebase. *Nucleic Acids Research*, 46(D1):D649–D655.

Hou, J. P. and Ma, J. (2013). *Identifying Driver Mutations in Cancer*, pages 33–56. Springer Netherlands.

Kim, Y.-A., Madan, S., and Przytycka, T. M. (2017). Wesme: uncovering mutual exclusivity of cancer drivers and beyond. *Bioinformatics*, 33(6):814–821.

Leiserson, M. D. M., Blokh, D., Sharan, R., and Raphael, B. J. (2013). Simultaneous identification of multiple driver pathways in cancer. *PLOS Computational Biology*, 9(5):1–15.

Leiserson, M. D. M., Vandin, F., Wu, H.-T., Dobson, J. R., Eldridge, J. V., Thomas, J. L., Papoutsaki, A., Kim, Y., Niu, B., McLellan, M., Lawrence, M. S., Gonzalez-Perez, A., Tamborero, D., Cheng, Y., Ryslik, G. A., Lopez-Bigas, N., Getz, G., Ding, L., and Raphael, B. J. (2015). Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nature Genetics*, 47(2):106–114.

Raphael, B. J., Dobson, J. R., Oesper, L., and Vandin, F. (2014). Identifying driver mutations in sequenced cancer genomes: computational approaches to enable precision medicine. *Genome Medicine*, 6(1):5.

Reyna, M. A., Leiserson, M. D. M., and Raphael, B. J. (2018). Hierarchical hotnet: identifying hierarchies of altered subnetworks. *Bioinformatics*, 34(17):i972–i980.

Stratton, M. R. (2009). The cancer genome. *Nature*, 458(7239):719–724.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., and Mesirov, J. P. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43):15545–15550.

Vandin, F., Upfal, E., and Raphael, B. J. (2011). Algorithms for detecting significantly mutated pathways in cancer. *Journal of Computational Biology*, 18(3):507–522.

Vandin, F., Upfal, E., and Raphael, B. J. (2012). De novo discovery of mutated driver pathways in cancer. *Genome research*, 22(2):375–385.