Randomisation of next generation sequencing data while preserving genomic event distributions

A. Gobbi[*1,4], **F. Iorio**[†2,3] **K.J. Dawson**[3], **D.C. Wedge**[3],
L.B. Alexandrov[3], **M.J. Garnett**[3], **G. Jurman**[‡4], **J. Saez-Rodriguez**[2]
[1] University of Trento – Trento (I)
[2] EMBL European Bioinformatics Institute – Cambridge (UK)
[3] Cancer Genome Project, Wellcome Trust Sanger Institute - Cambridge (UK)
[4] Fondazione Bruno Kessler – Trento (I)

Abstract

Studying combinatorial properties among patterns of mutations in cancer genomic datasets has recently emerged as a tool for identifying novel cancer hallmark, *e.g.*, the tendency of a set of genes to be mutated in a mutually exclusive manner. To this end, a Monte Carlo method (the switching-algorithm) is used for sampling simulated datasets under a null-model preserving patient- and gene-wise mutation rates. In this method, a genomic dataset is represented as a bipartite network, to which Markov chain updates (switching-steps) are applied to modify the network topology. A minimal number of them must be executed in order for the average similarity between the initial dataset and the randomised one to approximate that between two datasets generated independently under the null model. This number has been deducted empirically as a linear function of the total number of variants, making this process computationally expensive. Here we analytically derive an approximate lower bound for the number of steps required by the switching-algorithm. We also present BiRewire, an R/BioConductor package we developed by illustrating its use in generating 100,000 randomised versions of a breast cancer dataset and reporting a vast reduction in time requirement (from months to hours), with respect to existing implementations/bounds.

1 Introduction

Due to recent progress of Next Generation Sequencing (NGS) technologies, comprehensive catalogues of mutations in multiple cancer types have been assembled and fruitfully used to identify new diagnostic, prognostic and therapeutic targets. Existing large scale projects (such as the Cancer Genome Atlas - TCGA [1]) provide invaluable opportunities to explore molecular alterations that could potentially play a crucial role in a plethora of different cancer types and their response to therapy. A key task in these projects is to distinguish between driver mutations conferring selective clonal growth advantage and functionally neutral passenger mutations which do not contribute to tumour development. Considering then the context of the pathways where these key driver mutated genes operate allows the identification of 'cancer driver biological networks', whose altered functionality results in the acquisition of a 'cancer hallmark' [2]. One of the ideas exploited to identify these networks is based on the assumption that sets of mutations exhibiting statistically significant levels of mutual exclusivity are likely to alter genes involved in a common biological process that drives cancer development. Hence, driver mutations in cancer occur in a limited number of pathways and lesions in the same pathway do not tend to occur in the same patient. A possible biological

[*] Equally contributing authors
[†] Equally contributing authors
[‡] Corresponding author: Giuseppe Jurman `jurman@fbk.eu`

explanation is that if a crucial node is altered in an oncogenic pathway, a secondary mutation on the same pathway is unlikely to provide further selective advantages to the cancer cell, thus it does not tend to be evolutionary selected. On the other hand, mutations of genes participating in different biological pathways may exert a synergistic effect in conferring growth advantages to tumour cells. As a consequence, the combinatorial effects of gene mutations may play key roles in cancer initiation and progression, and the emergence of combinatorial properties among patterns of genomic events has been investigated in a number of recent studies, through the application of novel statistical measures quantifying, for example, the mutual exclusivity (ME) or the co-occurrence of different genomic lesions [3, 4, 5, 6].

In particular, Ciriello *et al.* [5] designed MEMo, a computational framework in which gene sets to be tested for ME are derived from cliques (*i.e.*, groups of genes pair-wisely connected) identified in functional networks, assembled from publicly available signaling- and pathway-maps. Compared to alternative methods, the functional relations occurring among a set of mutual-exclusive genes outputted by MEMo are more easily interpretable and the considered null model described hereafter reflects more comprehensively the statistical properties of the analysed genomic dataset. MEMo quantifies the sample coverage (SC) of a set of genes in terms of the number of samples in which at least one of them is mutated. Then the ME of the gene set under consideration is computed as the divergence of its SC from expectation. To this aim a null model is generated by randomly permuting the analysed dataset, while preserving the overall distribution of observed alterations across both genes and samples. This is crucial to preserve tumour specific alterations, heterogeneity in mutation/copy-number-alteration rates across patients, and to let the SC significance be proportional to the gene set ME. In order to generate this null model, the authors make use of a permutation strategy based on a random network generation model referred as the Switching-Algorithm (SA) [7]. Empirical p-values are then generated to estimate the significance of the deviation of the observed SC of each gene set from this null model.

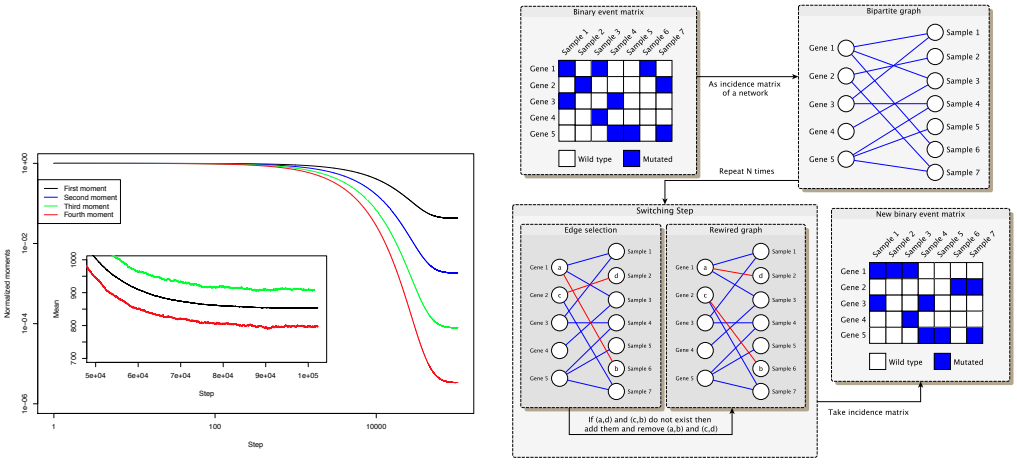
In practice, the analysed dataset is modeled as a Binary Event Matrix (BEM, see Fig. 1(b)), a '0-1 table' in which the generic entry is equal to 1 if a given gene is altered in a given sample, and 0 otherwise. In MEMo, randomised versions of a BEM are generated by adapting the switching-algorithm proposed in [7] to bipartite networks. This method uses a Markov chain and proceeds through a series of Monte Carlo Switching-Steps (SS) to produce rewired networks, starting from the original one, and preserving its degree distribution. The exact number of switching-steps that should be performed by the switching-algorithm in order to guarantee a proper level of mixing of the original network is not known. The authors propose on empirical grounds that 100 times the number of existing links is adequate, and this lower bound is generally used. In what follows we will refer to this bound as the empirical bound (N'). The desired number of random networks needed to estimate a null model and compute empirical p-values should then multiply this number, requiring a high computational cost.

Here we propose a novel, analytically derived, approximate lower bound, N, to the number of switching-steps required by the switching-algorithm in order to generate randomised versions of a bipartite graph, preserving genomic event distributions both across samples and genes, and achieving the maximum level of randomness. We have implemented BiRewire, an R/Bioconductor package allowing users: (i) to study and visualise trends of randomness across different number of switching-steps for a given BEM; (ii) to determine the minimum number of switching-steps required to maximize randomness while preserving genomic event distributions; (iii) to generate randomised BEMs in agreement with either this lower bound or a user-defined one. We illustrate the application of BiRewire with an example, and compare execution times on different implementations and bounds for the switching-algorithm, on a BEM derived from a TCGA breast cancer dataset, after the application of state-of-the-art filters for the identification of somatic mutations affecting protein function.

2 Results

We analytically derived a lower bound for the number of switching-steps to be performed by the switching-algorithm, when applied to a bipartite networks $\mathcal{G} = (V, E)$ (where V is the set of vertices and E the set of links, with $V = V_r \cup V_c$), in order to maximise the level of randomness in the resulting rewired network. This bound is equal to

$$N = \frac{e}{2(1-d)} \ln((1-d)e),$$



(a) (b)

Figure 1: (a) The first five normalised moments of the distribution of $x^{(k)}$, with the trend of the mean value and relative CI in the inset. (b) BEM randomisation, preserving genomic event distributions through the SA.

where e is the number of edges in the original network while d is its edge density, defined as the ratio between e and the number of edges in a fully connected bipartite graph with the same number of nodes in the two classes: $d = e(|V_r| \cdot |V_c|)$. If we count only successful switching-steps this bound is equal to $\frac{e(1-d)}{2} \ln((1-d)e)$ useful when the graph is dense and the degree distribution is far to be considered uniform. The starting point of our proof is the definition of similarity between a bipartite network $\mathcal{G}$ and its rewired version $\mathcal{G}^{(k+1)}$ based on the Jaccard Index $s^{(k)} = \frac{x^{(k)}}{2e - x^{(k)}}$ where $x^{(k)}$ is the bitwise sum of the Hadamart product between the incidence matrix of $\mathcal{G}$ and G(k), quantifying the number of edges in common between the original network and its rewired version after k SSs. The mean-field equation for $x^{(k+1)}$ is equal to:

$$x^{(k+1)} = \sum_{i=1}^{5} p_i^{(k)} f_i(x^{(k)})$$

where the functions $f_i(x^{(k)})$ represent five possible values of $x^{(k+1)}$ given $x^{(k)}$, depending on the switching step performing successfully or not and $p_i^{(k)}$ are the probabilities associated with these values. Specifying these probabilities allow the previous mean-field equation to be written as a second-order-linear recursive sequence $x^{(k+1)} = (m+1)x^{(k)} - mx^{(k-1)}$ for which a closed form is computable. More specifically,

$$x^{(k)} = m^k \left(e - \frac{q}{1-m} \right) + \frac{q}{1-m},$$

from which it is easy to find the unique fized point and the number of required SSs. With a similar procedure, a mean-field equation can also be estimated for the similarity between any pair of networks derived from the original network $\mathcal{G}$ through two different instances of the switching-algorithm, performing k switching-steps obtaining that the average similarity between any two rewired versions of a network $\mathcal{G}$ cannot be greater than the similarity between $\mathcal{G}$ and each of the two individual rewired versions. Formally showing that N switching-steps can simulate sampling from the uniform distribution of all the possible networks (with same nodes, number of edges and degree distributions of the starting one) would require a proof of convergence for the Markov chain underlying the switching-algorithm. Existence and uniqueness of a stationary distribution π for this process are guaranteed by this Markov chain being homogenous irreducible and with discrete-state. Nevertheless, even if it is relatively easy to derive the transition matrix of this process, it is not possible to compute a closed form for the $x^{(k)}$ distribution. For this reason we conducted an empirical study by simulating 2500 independent runs of the SA on an incidence matrix modeling

Table 1: Execution time for 100000 runs of the SA on the TCGA breast cancer data. With * we indicate extimated time.

	BiRewire (v. 0.99)	*igraph (v. 0.6.1)*	*igraph (v. 0.6.5)*
N	53m 20s	2d 2h 9m 36s	1y 67d 15h 34m 41s *
N'	9h 37m 30s	43d 4h 19m 12s	23y 355d 6h 52m *

a bipartite network with $n_c = 500, n_r = 1000$, and an edge density $d = 4\%$. We considered as a reference stationary distribution of the $x^{(k)}$ values the one reached after $N' = 100e$ switching steps (*i.e.*, the empirical bound proposed by Milo *et al.* [7]). Results of this simulation are depicted in figure 1, where we can see that the first five moments of the distribution of the $x^{(k)}$ tend to converge after N SS, together with the trend of the mean value of $x^{(k)}$ and its relative CI displayed in the inset. Together with our formal proof, these results show that N can be considered as an approximation of the convergence time (in terms of switching-steps) for the average level of edge-mixing of the original network. Moreover, they indicate that this property holds also for the whole distribution of the Markov chain underlying the switching-algorithm. The results shown above depends on the implementation of the SA. For this reason we developed an R package (submitted to Bioconductor and freely available at `http://www.bioconductor.org/packages/release/bioc/html/BiRewire.html`) providing high-performing routines for the generation of random bipartite graphs with prescribed node degrees through the switching-algorithm, for the analysis of randomness trends across SSs, and the estimation of the minimal number of steps.

As a benchmark datataset we used the TCGA breast cancer data, a 757×9757 binary matrix with 19758 not null entry ($d = 0.26\%$): in this case $N = 97951$ and $N' = 1975800 \sim 20N$. As shown in Tab.1, BiRewire is significantly faster than existing implementations, not only because it uses the newly detected lower bound, but also because it implements an optimal version of the switching-algorithm.

3 Conclusion

We analytically derived a novel lower bound for the minimal number of steps required by the switching-algorithm to randomise high-throughput genomic dataset, considerably reducing its computational time. We showed that this novel bound reduces the computational time requirements from months to hours, when tested on a real dataset and a typical desktop computer-architecture paired with an R package we developed. Our results can be generally adapted to the randomisation of any kind of dataset that can be modeled as a presence-absence matrix (hence a bipartite network) preserving the presence-distributions both across rows and columns. We believe that its applicability range covers different fields of computational biology and will grow in the future, as increasingly more data for which bipartite graphs provide a natural representation become available. Finally, we obtained a similar result (not shown here) for generic undirected networks.

References

[1] Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, 455(7216):1061–1068, 2008.

[2] D. Hanahan and R.A. Weinberg. Hallmarks of cancer: the next generation. *Cell*, 144(5):646–674, 2011.

[3] C. H. Yeang, F. McCormick, and A. Levine. Combinatorial patterns of somatic gene mutations in cancer. *The FASEB Journal*, 22(8):2605–2622, 2008.

[4] C. A. Miller, S. H. Settle, E. P. Sulman, K. D. Aldape, and A. Milosavljevic. Discovering functional modules by identifying recurrent and mutually exclusive mutational patterns in tumors. *BMC Medical Genomics*, 4(1):34, 2011.

[5] G. Ciriello, E. Cerami, C. Sander, and N. Schultz. Mutual exclusivity analysis identifies oncogenic network modules. *Genome Research*, 22(2):398–406, 2012.

[6] Y. Gu, D. Yang, J. Zou, W. Ma, R. Wu, W. Zhao, Y. Zhang, H. Xiao, X. Gong, M. Zhang, J. Zhu, and Z. Guo. Systematic Interpretation of Comutated Genes in Large-Scale Cancer Mutation Profiles. *Molecular Cancer Therapeutics*, 9(8):2186–2195, 2010.

[7] R. Milo, N. Kashtan, S. Itzkovitz, M.E.J. Newman, and U. Alon. On the uniform generation of random graphs with prescribed degree sequences. arXiv:cond-mat/0312028v2 [cond-mat.stat-mech], 2004.