Learning Cancer Progression Network from Mutation Allele Frequencies

Mohammad Sadegh Akhondzadeh [1] **Alireza Omidi** [2] **Zeinab Maleki** [1] **Kevin Coombes** [3] **Amanda E. Toland** [4] **Amir Asiaee** [5]

Abstract

We model the partial order of accumulation of mutations during tumorigenesis by linear structural equations. In this framework, the cancer progression network is modeled as a weighted directed acyclic graph (DAG), which minimizes a suitable continuous loss function. The goal is to learn the DAG from cross-sectional mutation allele frequency data. As a case study, we infer the order of mutations in melanoma. The recovered network of melanoma matches the known biological facts about the subtypes and progression of melanoma while discovers mutual exclusivity patterns among mutations by negative edges. Code implementing the proposed approach is open-source and publicly available at https://github.com/alirezaomidi/cancerdag.

1. Introduction

Cancer is a genetic disease where the accumulation of somatic alterations with selective advantage evolves the normal cells to a tumor. Learning this evolutionary process is critical for effective cancer treatment. Understanding the order of alterations leading to tumorigenesis is among the early objectives of cancer researchers (Vogelstein et al., 1988). The order in which alterations accumulate in the tumor cell population has shown to have clinical value (Beerenwinkel et al., 2015), helps in the refined staging of cancer (Vogelstein et al., 1988), and predicts the potential course of the disease (Hosseini et al., 2019).

[1]Department of Electrical and Computer Engineering, Isfahan University of Technology, Isfahan, Iran [2]Computer Engineering Department, Sharif University of Technology, Tehran, Iran [3]Department of Biomedical Informatics, Ohio State University, Columbus, USA [4]Department of Cancer Biology and Genetics and Department of Internal Medicine, Division of Human Genetics, Comprehensive Cancer Center, Ohio State University, Columbus, USA [5]Mathematical Biosciences Institute, Ohio State University, Columbus, USA. Correspondence to: Amir Asiaee <asiaeetaheri.1@osu.edu>.

Presented at the ICML 2020 Workshop on Computational Biology (WCB).

The main challenge in inferring the order of alterations is the fact that most large scale cancer data sets have low resolution and are cross-sectional (i.e., one sample at the time of diagnosis). Although the development of single-cell sequencing technologies is accelerating, the amount of single-cell data available from various cancers compared to bulk sequencing results from consortiums like TCGA (Cancer Genome Atlas Research Network et al., 2013) is minuscule. Besides, preprocessing and working with single-cell resolution data has its own computational and biological challenges. Issues like missing data (dropouts) and errors and biases due to whole genome amplification are complicating analysis and modeling of tumors using single-cell data (Lähnemann et al., 2020). In addition, the available data (bulk or single-cell) are usually sampled from a spatially heterogeneous tumor at the time of diagnosis (Marusyk et al., 2020) and therefore make it impossible to use phylogenetic reconstruction methods to infer the order of events. Thus, the goal of inferring the order of alterations is usually defined at the population level, i.e., we are interested in how the disease progresses on average in patients, which is recoverable using cross-sectional data from many tumors.

Cancer progression modeling arguably started with the work of Fearon and Vogelstein (Vogelstein et al., 1988), where they used data from precancerous lesions and tumors in various stages to build a chain progression model of colon cancer. Since then, many attempts have been made to generalize and extend cancer progression models. Modeling progression as trees (Desper et al., 1999), a mixture of trees (Beerenwinkel et al., 2005), and Directed Acyclic Graphs (DAGs) (Beerenwinkel et al., 2007) and learning the corresponding structures are among well-studied methods for inferring the order of alterations. More recent works consist of methods that utilize causality (Ramazzotti et al., 2015), pathway information (Gerstung et al., 2011), and flexible progression modeling (Nicol et al., 2020) to reconstruct more biologically plausible progression networks. Most of these methods have difficulties learning the mutual exclusivity relations of alterations present in tumors (Cristea et al., 2017) and need to take extra steps to find and incorporate those relations into their models (Ramazzotti et al., 2015; Nicol et al., 2020).

One of the main drawbacks of all of the papers in this do-

main is that they reduce the measurements to binary values. For example, for mutations, the measured values are mutation allele frequencies, but for ease of computation, the observed fractions are converted to zero one values. We believe that, especially in bulk sequencing data, this conversion results in a huge loss of information, which may end up being crucial in inferring the order of mutations.

In this work, we use linear structural equations to model the cancer progression DAG and propose a novel method for learning the DAG structure from mutation allele frequencies (MAFs). Our contributions are as follows:

- **Learning cancer progression from continuous mutation allele frequency data.** All of the previous models use a threshold to convert the continuous values of the mutation allele frequency to binary. Discarding MAFs and learning the progression network from binary data may eliminate relevant information contained in MAFs. To the best of our knowledge, we are the first to leverage MAFs in learning progression networks.

- **Capturing mutual exclusivity patterns between alterations.** It is known that mutations in genes of the same pathway are often mutually exclusive, because single perturbation of a pathway is sufficient for progression. For example, in melanoma, NRAS and BRAF mutations are seldom observed in the same samples. In contrast to state-of-the-art methods where the mutual exclusivity patterns should be learned separately and get injected into the progression network inference procedure, our proposed method learns them naturally while inferring the progression DAG.

2. Method

Consider n samples for which allele frequencies of d mutations $(X_1, \ldots, X_d)$ are measured. Here, we assume that there is an underlying DAG that represents the partial order under which mutations occur in a specific cancer type. We represent the observations by data matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$ where $x_{ij} \in [0, 1]$ are MAFs. We will discover a weighted DAG G with d nodes, which both describes the partial order of mutations during the tumorigenesis and mutual exclusivity relations. Each edge (i, j) has a weight $w_{ij} \in [-1, 1]$ where $w_{ij} \geq 0$ represents the conditional probability $\mathbb{P}(X_j = 1 | X_i = 1) = w_{ij}$ and $w_{ij} < 0$ indicates the mutual exclusivity of i and j, i.e., $\mathbb{P}(X_j = 0 | X_i = 1) = -w_{ij}$. So, the positive weights correspond to the partial order of progression and negative weights indicate mutual exclusivity relations of mutations.

2.1. Progression Model

We assume that each cell samples a path $p = (X_{p_1}, \ldots, X_{p_m})$ from the progression DAG, i.e., the cell state starts from normal and probabilistically accumulates mutations in the order dictated by the sampled path. Note that at step $X_{p_i} \rightarrow X_{p_{i+1}}$ the progression can stop with probability $1 - w_{p_i p_{(i+1)}}$. Besides the regular progression rule described above, there is a non-zero chance that mutation X_i can occur. Under this cell-wise progression model, one can write the number of cells with mutation i as:

$$N_i = \sum_{j=0}^{d} w_{ij} N_j + \epsilon_i, \quad \epsilon_i \sim N(0, N^2 \sigma^2) \tag{1}$$

where N_i is a random variable representing the number of cells with mutation i and ϵ_i models cells that has gain mutation i without following the given order (spontaneous activation describe in (Nicol et al., 2020)) or resisted acquiring mutation i despite the dictating order. Note that negative weights w_{ij} reduces N_i which is the correct behavior that we expect from mutual exclusive mutations i and j.

The causal relationship (1) between the number of cells with specific mutations is a form of structural equation model (Zheng et al., 2018). Dividing both sides by the total number of cells N, we get the following relationship between MAFs:

$$X_i = \sum_{j=0}^{d} w_{ij} X_j + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2) \tag{2}$$

2.2. Structure Learning

To learn W, often, we are interested in minimizing a loss $\mathcal{L}(W)$ subject to the DAGness constraint of the graph corresponding to W . The loss function is usually mean squared error penalized with ℓ_1 penalty to induce sparsity on W. Therefore the optimization becomes:

$$\min_{W} \quad \frac{1}{2n} \|X - XW\|_F^2 + \lambda \|W\|_1 \quad \text{s.t.} \qquad G \in \mathbb{D} \tag{3}$$

where $\|.\|_F$ is the Frobenius norm, $\|.\|_1$ is ℓ_1 norm, λ is the penalty coefficient, and $\mathbb{D}$ is the discrete set of all possible DAGs with d nodes.

The main issue with the optimization problem (3) is its combinatorial nature due to the $G \in \mathbb{D}$ constraint. The set $\mathbb{D}$ growths superexponentially in d and makes solving the problem exactly NP-hard. Recently, the following continuous measure was introduced in (Zheng et al., 2018) for the characterization of acyclicity.

Theorem 1 *A matrix $W \in \mathbb{R}^{d \times d}$ corresponds to a DAG if and only if*

$$h(W) = tr(e^{W \circ W}) - d = 0 \tag{4}$$

where $\circ$ is the Hadamard product operator. Moreover, $h(W)$ has a simple gradient $\nabla h(W) = \left(e^{W \circ W}\right)^T \circ 2W$.

Note that $h(W) \geq 0$ is a smooth differentiable continuous function of a weight matrix W, whose value indicates "DAG-ness" of G. In other words, $h(W) = 0$ for DAGs and in loopy graphs, as the weight of loops decreases, $h(W)$ becomes smaller. Next, (Zheng et al., 2018) suggest to solve the following non-convex problem (the constraint set is non-convex) by the augmented Lagrangian method:

$$\min_W \frac{1}{2n} \|X - XW\|_F^2 + \lambda \|W\|_1 \text{ s.t. } h(W) = 0 \quad (5)$$

For the cancer progression inference problem described in Section 2.1, we solve the exact objective of (5) for a stationary point with an extra box constraint for elements of W, i.e., $w_{ij} \in [-1, 1]$. Finally, at the end, we adopt two different threshold for positive and negative edges to shrink non confident edges to zero.

3. Results

One of the major issues in validating the inferred progression network is the lack of ground truth biological knowledge about the progression of cancer. We will consider melanoma for evaluating our proposed method because we have a partial biological understanding of its progression. In fact, because of frequent screening for melanoma, precancerous lesions and tumor samples are obtained in different stages across the patient population. Thus, we have a better understanding of early and late alterations (Shain et al., 2015; Akbani et al., 2015). Here, we apply our proposed method to metastatic melanoma data of The Cancer Genome Atlas (**TCGA**) (Cancer Genome Atlas Research Network et al., 2013) and show that we can recover the known biological facts about the progression of melanoma.

3.1. Data and Preprocessing

The Mutation Allele Frequency (MAF) data of TCGA's Skin Cutaneous Melanoma (SKCM) is downloaded from the NCI Genomic Data Commons Data Portal. The data matrix consists of 470 samples and 21220 features, which are MAFs of all genes. As a preprocessing step, a set of driver mutations are selected using the results of a recent study (Bailey et al., 2018), where multiple driver-gene identification methods and tools have been combined and tested on TCGA's data. The reported driver genes of melanoma in (Bailey et al., 2018) consists of 24 different genes. Next, we filtered out some of the genes with the insights provided by cBioPortal tool (Cerami et al., 2012; Gao et al., 2013). The discarded genes have no reported mutation hotspots, have a small proportion of nonsense mutations (truncating), and their reported missense mutations are uniformly spread across their cDNA with no known pathogenicity. These properties suggest that the discarded mutations tend to have no significant effects in melanoma. The preprocessing step leaves us with $n = 439$ samples and $d = 19$ driver-genes.

3.2. Recovered Progression Network

To quantify the uncertainty of recovered edge weights, we use bootstrap. We run our method on 100 bootstrap samples and report the average weight for each edge. We then use the t-test to check that the average percentage of times an edge is recovered is over the given threshold of τ. For this experiment, we pruned positive edges of weights below the threshold of $w_+ = 0.09$ and negative edges of weights above the threshold of $w_- = -0.01$.

The inferred progression DAG of melanoma is illustrated in Figure 1. Note that the final number of nodes in the figure is 15 because four of the driver mutations (*KIT*, *GNA11*, *HRAS* and *KRAS*) were isolated and therefore we omitted them.

BRAF, *NRAS*, *NF1*, and *TP53* have the highest rate of mutation in melanoma. The negative edges of (*BRAF*, *NRAS*) and (*BRAF*, *NF1*) suggest that they are mutually exclusive, as reported before (Akbani et al., 2015). After removing the negative edges (in red), the remaining graph represents the progression network. The main roots of the network are *BRAF* and *NRAS* and they share many descendants where the most important one is *TP53*. Since *NF1* is not co-occurring with *BRAF* it seems that the progression order toward *NF1* is *NRAS* $\rightarrow$ *TP53* $\rightarrow$ *NF1*. The strongest observed edge in the graph is (*BRAF*, *PTEN*).

4. Discussion

There are various studies (Kunz, 2014; Akbani et al., 2015; Rajkumar & Watson, 2016) on skin cancer, attempting to distinguish melanoma subtypes. Based on these studies, there are four main subtypes for melanoma. The *BRAF* mutation identifies the largest genomic subtype. This subtype consists of more than half of the melanoma cases. Our recovered network places *BRAF* at one of the roots of progression, which is aligned with the fact that it defines a subtype and also has been known to occur early in melanoma (Shain et al., 2015). The *RAS* mutation family ($\{$*N,K,H*$\}$*-RAS* but mostly *NRAS*) determines the second genomic subtype of melanoma, which occurs in about a quarter of melanoma tumors. *NRAS* is selected as another root for the progression graph, which is in agreement with being a subtype hallmark.

The third subtype of melanoma is identified by *NF1* mutation. This mutation has happened in about 30% of samples, which has no *BRAF* or *NRAS* mutations. Our model does not capture the *NF1* as a separate root but recovers a negative edge between *BRAF* and *NF1* which suggests mutual exclusivity of the two. The last melanoma subtype is Triple-wild, i.e., samples without any of the *BRAF*, *NRAS*, and

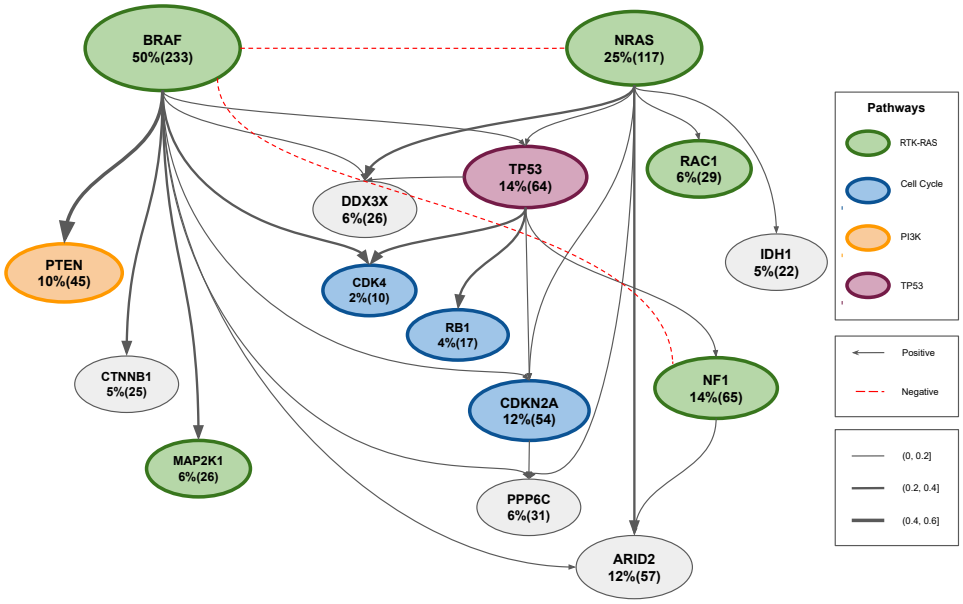

Figure 1. Inferred Progression Network of Melanoma

NF1 mutations. This subtype's frequency is much lower than the other ones (lower than 10%). The most related mutation attributed to this subtype occurs in *KIT*, which is present in around 4% of samples. *KIT* mutation becomes an isolated node and is not shown in the progression network of Figure 1.

Our method captures negative edges (*BRAF*, *NRAS*) and (*BRAF*, *NF1*), which suggests their mutual exclusivity. Mutual exclusivity of both pairs has been reported in the literature (Davies et al., 2002; Davies & Samuels, 2010). The recovered negative weights show our proposed model's ability to learn mutual exclusivity relations simultaneously along with the progression network.

There are studies which claim that mutation in *CDKN2A* and *TP53* occur in intermediate and advanced melanoma (Shain et al., 2015; Davis et al., 2018). In our recovered progression network, they occur after *BRAF* or *NRAS*. Finally, the *TP53* mutation is known to be a frequent mutation occurring in all major subtypes of *BRAF*, *NRAS*, and *NF1* (Davis et al., 2018). In our inferred progression DAG, there are direct paths that connect *TP53* and all major mutations of various subtypes.

5. Conclusion

In this paper, we presented a novel approach for inferring cancer progression network from mutation allele frequency of cross-sectional tumor data. We formulated the problem as a continuous non-convex optimization over the class of DAGs, representing the underlying partial order of mutations occurring in cancer. The proposed method is able to take allele frequencies as input and finds mutually exclusive mutations while learning the progression network. We demonstrated these abilities of the method by using the TCGA Skin Cutaneous Melanoma data set as a case study. We showed that our model recovers a progression network that matches the known biological facts about tumorigenesis of melanoma and also captures the well-established mutual exclusivity patterns that are genomic hallmarks of melanoma subtypes.

References

Akbani, R., Akdemir, K. C., Aksoy, B. A., Albert, M., Ally, A., Amin, S. B., Arachchi, H., Arora, A., Auman, J. T., Ayala, B., et al. Genomic classification of cutaneous melanoma. *Cell*, 161(7):1681–1696, 2015.

Bailey, M. H., Tokheim, C., Porta-Pardo, E., Sengupta, S., Bertrand, D., Weerasinghe, A., Colaprico, A., Wendl, M. C., Kim, J., Reardon, B., et al. Comprehensive characterization of cancer driver genes and mutations. *Cell*, 173(2):371–385, 2018.

Beerenwinkel, N., Rahnenführer, J., Däumer, M., Hoffmann, D., Kaiser, R., Selbig, J., and Lengauer, T. Learning multiple evolutionary pathways from cross-sectional data. *Journal of computational biology*, 12(6):584–598, 2005.

Beerenwinkel, N., Eriksson, N., and Sturmfels, B. Conjunctive bayesian networks. *Bernoulli*, pp. 893–909, 2007.

Beerenwinkel, N., Schwarz, R. F., Gerstung, M., and Markowetz, F. Cancer evolution: mathematical models and computational inference. *Systematic biology*, 64 (1):e1–25, January 2015.

Cancer Genome Atlas Research Network, Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R. M., Ozenberger, B. A., Ellrott, K., Shmulevich, I., Sander, C., and Stuart, J. M. The cancer genome atlas Pan-Cancer analysis project. *Nature genetics*, 45(10):1113–1120, October 2013.

Cerami, E., Gao, J., Dogrusoz, U., Gross, B. E., Sumer, S. O., Aksoy, B. A., Jacobsen, A., Byrne, C. J., Heuer, M. L., Larsson, E., et al. The cbio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data, 2012.

Cristea, S., Kuipers, J., and Beerenwinkel, N. pathTiMEx: Joint inference of mutually exclusive cancer pathways and their progression dynamics. *Journal of computational biology: a journal of computational molecular cell biology*, 24(6):603–615, June 2017.

Davies, H., Bignell, G. R., Cox, C., Stephens, P., Edkins, S., Clegg, S., Teague, J., Woffendin, H., Garnett, M. J., Bottomley, W., et al. Mutations of the braf gene in human cancer. *Nature*, 417(6892):949–954, 2002.

Davies, M. A. and Samuels, Y. Analysis of the genome to personalize therapy for melanoma. *Oncogene*, 29(41): 5545–5555, 2010.

Davis, E. J., Johnson, D. B., Sosman, J. A., and Chandra, S. Melanoma: What do all the mutations mean? *Cancer*, 124(17):3490–3499, 2018.

Desper, R., Jiang, F., Kallioniemi, O.-P., Moch, H., Papadimitriou, C. H., and Schäffer, A. A. Inferring tree models for oncogenesis from comparative genome hybridization data. *Journal of computational biology*, 6(1): 37–51, 1999.

Gao, J., Aksoy, B. A., Dogrusoz, U., Dresdner, G., Gross, B., Sumer, S. O., Sun, Y., Jacobsen, A., Sinha, R., Larsson, E., et al. Integrative analysis of complex cancer genomics and clinical profiles using the cbioportal. *Sci. Signal.*, 6 (269):pl1–pl1, 2013.

Gerstung, M., Eriksson, N., Lin, J., Vogelstein, B., and Beerenwinkel, N. The temporal order of genetic and pathway alterations in tumorigenesis. *PloS one*, 6(11): e27136, November 2011.

Hosseini, S.-R., Diaz-Uriarte, R., Markowetz, F., and Beerenwinkel, N. Estimating the predictability of cancer evolution. *Bioinformatics*, 35(14):i389–i397, July 2019.

Kunz, M. Oncogenes in melanoma: an update. *European journal of cell biology*, 93(1-2):1–10, 2014.

Lähnemann, D., Köster, J., Szczurek, E., McCarthy, D. J., Hicks, S. C., Robinson, M. D., Vallejos, C. A., Campbell, K. R., Beerenwinkel, N., Mahfouz, A., Pinello, L., Skums, P., Stamatakis, A., Attolini, C. S.-O., Aparicio, S., Baaijens, J., Balvert, M., Barbanson, B. d., Cappuccio, A., Corleone, G., Dutilh, B. E., Florescu, M., Guryev, V., Holmer, R., Jahn, K., Lobo, T. J., Keizer, E. M., Khatri, I., Kielbasa, S. M., Korbel, J. O., Kozlov, A. M., Kuo, T.-H., Lelieveldt, B. P. F., Mandoiu, I. I., Marioni, J. C., Marschall, T., Mölder, F., Niknejad, A., Raczkowski, L., Reinders, M., Ridder, J. d., Saliba, A.-E., Somarakis, A., Stegle, O., Theis, F. J., Yang, H., Zelikovsky, A., McHardy, A. C., Raphael, B. J., Shah, S. P., and Schönhuth, A. Eleven grand challenges in single-cell data science. *Genome biology*, 21(1):31, February 2020.

Marusyk, A., Janiszewska, M., and Polyak, K. Intratumor heterogeneity: The rosetta stone of therapy resistance. *Cancer cell*, 37(4):471–484, April 2020.

Nicol, P. B., Coombes, K. R., Deaver, C., Chkrebtii, O. A., Paul, S., Toland, A. E., and Asiaee, A. Oncogenetic network estimation with disjunctive bayesian networks: Learning from unstratified samples while preserving mutual exclusivity relations. *bioRxiv*, 2020.

Rajkumar, S. and Watson, I. R. Molecular characterisation of cutaneous melanoma: creating a framework for targeted and immune therapies. *British journal of cancer*, 115(2):145–155, 2016.

Ramazzotti, D., Caravagna, G., Olde Loohuis, L., Graudenzi, A., Korsunsky, I., Mauri, G., Antoniotti, M., and Mishra, B. Capri: efficient inference of cancer progression models from cross-sectional data. *Bioinformatics*, 31(18):3016–3026, 2015.

Shain, A. H., Yeh, I., Kovalyshyn, I., Sriharan, A., Talevich, E., Gagnon, A., Dummer, R., North, J., Pincus, L., Ruben, B., et al. The genetic evolution of melanoma from precursor lesions. *New England Journal of Medicine*, 373(20): 1926–1936, 2015.

Vogelstein, B., Fearon, E. R., Hamilton, S. R., Kern, S. E., Preisinger, A. C., Leppert, M., Smits, A. M., and Bos, J. L. Genetic alterations during colorectal-tumor development. *New England Journal of Medicine*, 319(9):525–532, 1988.

Zheng, X., Aragam, B., Ravikumar, P. K., and Xing, E. P. Dags with no tears: Continuous optimization for structure learning. In *Advances in Neural Information Processing Systems*, pp. 9472–9483, 2018.