PERFECT RECONSTRUCTION OF ONCOGENETIC TREES

CHARALAMPOS E. TSOURAKAKIS

Abstract. In this note we provide the necessary and sufficient conditions to uniquely reconstruct an oncogenetic tree.

1. Introduction

Human cancer is caused by the accumulation of genetic alternations in cells [1, 19]. Finding driver genetic mutations, i.e., mutations which confer growth advantage on the cells carrying them and have been positively selected during the evolution of the cancer and uncovering their temporal sequence have been central goals of cancer research the last decades [16]. Among the triumphs of cancer research stands the breakthrough work of Vogelstein and his collaborators [8, 18] which provides significant insight into the evolution of colorectal cancer. Specifically, the so-called "Vogelgram" models colorectal tumorigenesis as a linear accumulation of certain genetic events. Few years later, Desper et al. [6] considered more general evolutionary models compared to the "Vogelgram" and presented one of the first theoretical approaches to the problem [1], the so-called *oncogenetic trees*. Before we provide a description of oncogenetic trees which are the focus of our work, we would like to emphasize that since then a lot of research work has followed from several groups of researchers, influenced by the seminal work of Desper et al. [6]. Currently there exists a wealth of methods that infer evolutionary models from microarray-based data such as gene expression and array Comparative Genome Hybridization (aCGH) data: distance based oncogenetic trees [7], maximum likelihood oncogenetic trees [11], hidden variable oncogenetic trees [17], conjuctive Bayesian networks [3] and their extensions [5, 9], mixture of trees [4]. The interested reader is urged to read the surveys of Attolini et al. [1] and Hainke et al. [10] and the references therein on established progression modeling methods. Furthermore, oncogenetic trees have successfully shed light into many types of cancer such as renal cancer [6], hepatic cancer [13] and head and neck squamous cell carcinomas [12].

Oncogenetic Trees. An oncogenetic tree is a rooted directed tree[1]. The root represents the state of tissue with no mutations. Any other vertex $v \in V$ represents a mutation. Each edge represents a "cause-and-effect" relationships. Specifically, for a mutation represented by vertex v to occur, all the mutations corresponding to vertices that lie on the directed path from the root to v must be present in the tumor. In other words, if two mutations u, v are connected by an edge (u, v) then v cannot occur if u has not occured. The edges are labeled with probabilities. Each tumor corresponds to a rooted subtree of the oncogenetic tree and the probability of occurence is determined as described by [6]. Desper et al. provide an algorithm that finds a likely oncogenetic tree that fits the observed data.

In this work we answer a fundamental question regarding oncogenetic trees. Before we state the question, we introduce some notation first. Let $T = (V, E, r)$ be a rooted tree on V, i.e., every

[1] Typically, the term *tree* is reserved for the undirected case and the term *branching* for the directed case. Throughtout this note, we consistently use the term *tree* mean a directed tree as in [6].

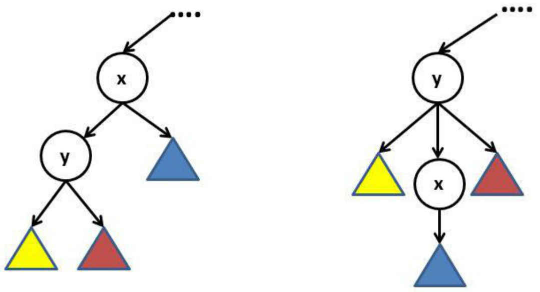	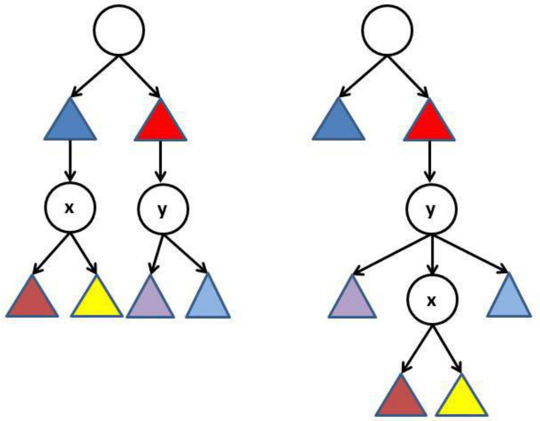
Figure 2(a) Necessity of Condition 1	Figure 2(b) Necessity of Condition 2

Table 1. Illustration of necessity conditions of Theorem 1.

vertex has in-degree at most one and there are no cycles, and let $r \in V$ be the root of T. Given a finite family $\mathcal{F} = \{A_1, ...A_q\}$ of sets of vertices, i.e., $A_i \subseteq V(T)$ for $i = 1, \ldots, q$, where each A_i is the vertex set of a r-rooted sub-tree of T, what are the necessary and sufficient conditions, if any, which allow us to uniquely reconstruct T? In this work we treat this natural combinatorial question, namely:

"When can we reconstruct an oncogenetic tree T from a set family $\mathcal{F}$?"

Despite the fact that in practice aCGH data tend to be noisy and consistent with more than one oncogenetic trees, the question is nonetheless interesting and to the best of our knowledge remains open so far [14, 15]. Theorem 1 provides the necessary and sufficient conditions to uniquely reconstruct an oncogenetic tree. We write $u \prec v$ ($u \not\prec v$) to denote that u is (not) a descendant of v in T.

Theorem 1. *Let T be an oncogenetic tree and $\mathcal{F} = \{A_1, ...A_q\}$ be a finite family of sets of vertices, i.e., $A_i \subseteq V(T)$ for $i = 1, \ldots, q$, where each A_i is the vertex set of a r-rooted sub-tree of T The necessary and sufficient conditions to uniquely reconstruct the tree T from the family $\mathcal{F}$ are the following:*

1. *For any two distinct vertices $x, y \in V(T)$ such that $(x, y) \in E(T)$, there exists a set $A_i \in \mathcal{F}$ such that $x \in A_i$ and $y \notin A_i$.*
2. *For any two distinct vertices $x, y \in V(T)$ such that $y \not\prec x$ and $x \not\prec y$ there exist sets $A_i, A_j \in \mathcal{F}$ such that $x \in A_i$, $y \notin A_i$ and $x \notin A_j$ and $y \in A_j$.*

In Section 2 we prove Theorem 1. It is worth noticing that our proof provides a simple procedure for the reconstruction as well.

2. Proofs

In the following we call a tree T *consistent* with the family set $\mathcal{F}$ if all sets $A_i \in \mathcal{F}$ are vertices of rooted sub-trees of T. Notice that when two or more trees are consistent with the input dataset $\mathcal{F}$, then we cannot uniquely reconstruct T.

Proof. First we prove the necessity of conditions 1,2 and then their sufficiency to reconstruct T.

Necessity: For the sake of contradiction, assume that Condition 1 does not hold. Therefore, there exists two vertices $x, y \in V(T)$ such that there exists no set $A \in \mathcal{F}$ that contains one of them. Then, the two trees shown in Figure 2(a) are both consistent with $\mathcal{F}$. Therefore we cannot reconstruct T. Similarly, assume that Condition 2 does not hold. Specifically assume that for all j such that $x \in A_j$, then $y \in A_j$ too (for the symmetric case the same argument holds). Then, both trees in Figure 2(b) are consistent with $\mathcal{F}$ and therefore T is not reconstructable. The symmetric case follows by the same argument.

Sufficiency: Let $x \in V(T)$ and P_x be the vertex set of the unique path from the root r to x, i.e., $P_x = \{r, \ldots, x\}$. Also, define F_x to be the intersection of all sets in the family $\mathcal{F}$ that contain vertex x, i.e., $F_x = \bigcap\limits_{A_i \ni x} A_i$. We prove that $F_x = P_x$. Since by the definition of an oncogenetic tree $P_x \subseteq F_x$ it suffices to show that $F_x \subseteq P_x$. Assume for the sake of contradiction that $F_x \nsubseteq P_x$. Then, there exists a vertex $v \in V(T)$ such that $v \in F_x, v \notin P_x$. We consider the following three cases.

- Case 1 $(x \prec v)$: Since by definition each set $A_i \in \mathcal{F}$ is the vertex set of a rooted sub-tree of T, $v \in P_x$ by the definition of an oncogenetic tree.
- Case 2 $(v \prec x)$: Inductively by condition 1, there exists $A_i \in \mathcal{F}$ such that $x \in A_i, v \notin A_i$. Therefore, $v \notin F_x$.
- Case 3 $(x \not\prec v, v \not\prec x)$: By condition 2, there exists $A_i \in \mathcal{F}$ such that $x \in A_i$ and $v \notin A_i$. Hence, $v \notin F_x$.

In all three cases above, we obtain a contradiction and therefore $v \in F_x \Rightarrow v \in P_x$. Therefore, $F_x \subseteq P_x$ and subsequently $F_x = P_x$. Given this fact, it is easy to reconstruct the tree T. We sketch the algorithm: compute for each x the set F_x which is the unordered set of vertices of the unique path from r to x. The ordering of the vertices which results in finding the path P_x, i.e., $(v_0 = r \to v_1 \to ..v_{k-1} \to v_k = x)$ is computed using sets in $\mathcal{F}$ which contain v_i but not v_{i+1}, $i = 0, .., k-1$. The existence of such sets is guaranteed by condition 1. □

3. Acknowledgments

Research supported by NSF Grant No. CCF-1013110. The author thanks Deepak Bal and Professor Alan Frieze for their feedback.

References

[1] Attolini, C. S.-O., Michor, F.: *Evolutionary Theory of Cancer.* Annals of the New York Academy of Sciences, Vol. 1168, pp. 2351 (2009)

[2] Beerenwinkel, N., Rahnenführer, J., Kaiser, R., et al.: *Mtreemix: a software package for learning and using mixture models of mutagenetic trees.* Bioinformatics, Vol. 21, pp. 21062107 (2005)

[3] Beerenwinkel, N., Eriksson, N., Strumfels, B.: *Conjunctive bayesian networks.* Bernoulli, Vol. 13, pp. 893909 (2007)

[4] Beerenwinkel, N., Rahnenführer, J., Däumer, M., Hoffmann, D., Kaiser, R., Selbig, J., Lengauer, T.: *Learning multiple evolutionary pathways from cross-sectional data.* Journal of Computational Biology, Vol. 12, pp. 584598 (2005)

[5] Beerenwinkel, N., Sullivant, S.: *Markov models for accumulating mutations.* Biometrika, Vol. 96, pp. 663676 (2009)

[6] Desper, R., Jiang, F., Kallioniemi, O.P., Moch, H., Papadimitriou, C.H., Schäffer, A.A.: *Inferring tree models for oncogenesis from comparative genome hybridization data.* Journal of Computational Biology, Vol. 6(1), pp. 37-51 (1999)

[7] Desper, R., Jiang, F., Kallioniemi, O.P., Moch, H., Papadimitriou, C.H., Schäffer, A.A.: *Distance-based reconstruction of tree models for oncogenesis.* J. Comput. Biol., Vol. 7(6), pp. 789-803 (2000)

[8] Fearon, E. R., Vogelstein, B.: *A genetic model for colorectal tumorigenesis.* Cell, Vol. 61, pp. 759767 (1990)

[9] Gerstung, M., Baudis, M., Moch, H., Beerenwinkel, N.: *Quantifying cancer progression with conjunctive bayesian networks.* Bioinformatics, Vol. 25, pp. 28092815 (2009)

[10] Hainke, K., Rahnenführer, J., Fried, R.: *Disease progression models: A review and comparison.* Dortmund University, Technical Report

[11] Heydebreck, A., Gunawan, B., Füzesi, L.: *Maximum likelihood estimation of oncogenetic tree models.* Biostatistics, Vol. 5(4), pp. 545556 (2004)

[12] Huang, Q., Yu, G.P., Mo, J., Datta, B., Mahimkar, M., Lazarus, P., Schäffer, A.A., Desper, R., Schantz, S. P.: *Genetic differences detected by comparative genomic hybridization in head and neck squamous cell carcinomas from different tumor sites: construction of oncogenetic trees for tumor progression.* Genes, Chromosomes and Cancer, Vol. 34(2), pp, 224233 (2002)

[13] Longerich, T., Mueller, M.M, Breuhagn, K., Schirmacher, P., Benner, A., Heiss, C.: *Oncogenetic tree modeling of human hepatocarcinogenesis.* International Journal of Cancer, Vol. 130(3), pp. 575583 (2012)

[14] Papadimitriou, C.: *Personal Communication*

[15] Schäffer, A.A.: *Personal Communication*

[16] Stratton M.R., Campbell, P.J., Futreal, P.A.: *The cancer genome.* Nature, Vol. 458(7239), pp. 719-724 (2009)

[17] Tofigh, A.: *Using trees to capture reticulate evolution: lateral gene transfers and cancer progression.* Ph.D. thesis (2009)

[18] Vogelstein, B., Fearon, E.R., Hamilton, S.R., Kern, S.E., Preisinger, A.C., Leppert, M., Nakamura, Y., White, R., Smits, A.M., Bos, J.L.: *Genetic alterations during colorectal-tumor development.* New England Journal of Medicine, Vol. 319, pp. 525532 (1988)

[19] Weinberg, R. A.: *The Biology of Cancer.* Garland Science (2007)

Department of Mathematical Sciences, Carnegie Mellon University, 5000 Forbes Av., 15213, Pittsburgh, PA, U.S.A

E-mail address: ctsourak@math.cmu.edu