Analysis of Zero-Agnostic Model for Copy Number Evolution

Sonali Merchia

1. Introduction

Reconstructing cancer phylogenies based on copy-number information is widely done using the copy number transformation CNT model. This model splits up chromosomes into geographical regions called loci. It then represents a cell using a vector where each element represents a locus and the value represents the number of copies of that locus in the cell of interest. The model then measures the distance between a pair of copy number profiles as the minimum number of mutations required to convert one state into the other [5].

While useful and consistent with biological observations, there is no efficient algorithm using CNT to generate parsimonious cancer phylogeneies due to its lack of symmetry. In *A zero-agnostic model for copy number evolution in cancer* [4], a new model ZCNT is proposed that provides a mathematical simplification on the CNT model, allowing for an efficient algorithm to generate cancer phylognies.

ZCNT allows for elements of the copy-number state to drop into the negatives and increase from zero to positive values. This is not possible in CNT because it would represent spontaneous creation of a locus that did not exist previously. This paper will analyze the ZCNT method to determine whether its findings are consistent with biology despite being based on a model inconsistent with biology.

2. Background

2.1. Cancer as an Evolutionary Process

Cancer is an evolutionary process wherein cells grow and multiply irregularly due to genetic mutations. Mutations are alterations of the DNA stored within cancer cells and take multiple different forms. The mutations are random and unique in each tumor but lead to the development of common hallmarks.

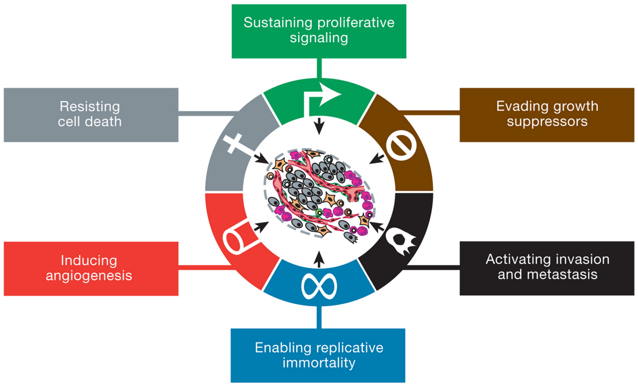
Sustaining proliferative signaling
Resisting cell death
Evading growth suppressors
Inducing angiogenesis
Activating invasion and metastasis
Enabling replicative immortality

Figure 1: Hallmarks of Cancer [3]

Since these hallmarks are acquired through evolution, studying the phylogeneies of cancer cells can help inform how the cells acquired their abilities, providing insight into tumor behaviors and opening new avenues for potential treatments.

2.2. CNA-Based Phylogenies

Different kinds of mutations affect the genome to different magnitudes. The most commonly studied kinds of mutations in cancer phylogenetics are Single Nucleotide Variants (SNVs) and Copy Number Aberrations (CNAs). SNVs are when a single base pair in the DNA is erroneously changed, inserted, or deleted. Meanwhile CNAs are when large geographical regions of base pairs are erroneously duplicated or deleted. The CNT and ZCNT models that are discussed in this paper will focus on CNAs.

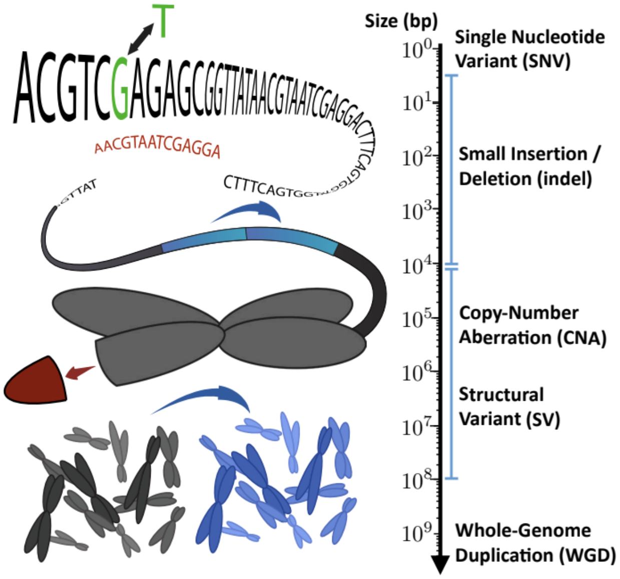
T
ACGTCGAGAGCGGTTATAACGTAATCGAGGACTTTCAGTGG
AACGTAATCGAGGA
TTAT
CTTTCAGTGG
Size (bp)
10^0
10^1
10^2
10^3
10^4
10^5
10^6
10^7
10^8
10^9
Single Nucleotide Variant (SNV)
Small Insertion / Deletion (indel)
Copy-Number Aberration (CNA)
Structural Variant (SV)
Whole-Genome Duplication (WGD)

Figure 2: Different kinds of mutations affect genomes differently and their affects are sometimes of vastly different scales. This graphic relates different kinds of mutations to the scale of base pairs affected by a mutation of that type. [1]

There are two main phylogenetic problems to be solved: the Large Parsimony Problem and the Small Parsimony Problem. Large Parsimony is when you are given the copy-number states of cells and aim to organize them into a rooted evolutionary tree wherein every cell has exactly one parent. Meanwhile Small Parsimony is when you are given the copy-number states of "leaf-node" cells and aim to generate the intermediary nodes of the phylogenetic tree.

The core principle behind cancer phylogeny reconstruction is that mutations are rare. That means that in order to reconstruct cancer phylogeneies, it is useful to define distance functions representative of the number of mutations. These allow us to construct phylogenetic trees that would require few mutations to generate, encouraging parsimonious evolutionary trees. Both ZCNT and CNT define different algorithms to determine distance.

2.3. CNA-Based Distance Functions

Both CNT and ZCNT models have mathematical definitions of events. In both models events represent a copy-number aberration that affects copy number profile p over a range of loci from s to t inclusive ($s \leq t$) by increasing or decreasing the copy number by 1 (denoted by $b \in \{+1, -1\}$).

They also both define the concept of a "transformation" to be a series of events. For instance, given events $(e_1, \ldots, e_n)$, a transformation would be defined as $T(p) = e_n(\ldots(e_1(p)))$. Both models define distance between two profiles to be the minimum number of events required to transform one profile into another ($\text{dist}(u, v) = \min_{T(u)=v} |T|$).

2.3.1 CNT

Copy Number Event Formula: ($c_{s,t,b} : \mathbb{Z}_+^n \to \mathbb{Z}_+^n$)

$$c_{s,t,b}(p)_i = \begin{cases} p_i + b & \text{if } s \leq i \leq t \text{ and } p_i \neq 0 \\ p_i & otherwise \end{cases} \tag{1}$$

If given two copy number profiles, it is not guaranteed that there be a transformation from one to the other. In this case the distance is defined as $+\infty$. We will denote CNT distance between two profiles u and v as $dist_{CNT}(u, v)$.

Due to the lack of symmetry, two common methods have been employed to create symmetry:

Method	**Formula**
Mean Correction	$\frac{dist_{CNT}(u,v)+dist_{CNT}(v,u)}{2}$
Median Distance	$\min dist_{CNT}(u, v) + dist_{CNT}(v, u)$

Both of these definitions raise interesting questions about reachability. Mean correction amplifies unreachability while median distance amplifies reachability.

Instead of using a distance with these constraints, it is common to use another kind of intermediary. Some papers use the *copy number triplet* (CN3) problem. Which aims to pick some intermediate profile w and minimize $dist_{CNT}(w, u) + dist_{CNT}(w, v)$ [2]. While CN3 provides symmetry, it comes at the cost of computational time. CNT distances can be computed in $O(n)$ time [6] while the most optimal current algorithm for CN3 is $O(nB^7)$ where B is the largest copy number in the input profiles.

2.3.2 ZCNT

Zero-Agnostic Copy Number Event Formula
($c_{s,t,b} : \mathbb{Z}^n \to \mathbb{Z}^n$)

$$c_{s,t,b}(p)_i = \begin{cases} p_i + b & \text{if } s \leq i \leq t \\ p_i & otherwise \end{cases} \tag{2}$$

Notice that the only difference between the ZCNT and CNT model is the removal of the non-negativity constraint in the function signature and the condition to skip p_i if p_i is zero. We will denote ZCNT distance between two profiles u and v as $dist_{ZCNT}(u, v)$.

In *A zero-agnostic model for copy number evolution in cancer* [4], they propose the following closed-form formula for ZCNT distance:

$$dist_{ZCNT}(u, v) = \frac{1}{2}||\Delta(u) - \Delta(v)||_1 \tag{3}$$

where $\Delta(p)$ denotes another linear transform ($\Delta : \mathbb{Z}^n \to \mathbb{Z}^{n+1}$)

$$\Delta(p)_i = \begin{cases} p_1 - 2 & \text{if } i = 1 \\ 2 - p_n & \text{else if } i = n + 1 \\ p_i - p_{i-1} & otherwise \end{cases} \tag{4}$$

Due to ZCNT being the composition of two linear equations (Equations 3 and 4), ZCNT can trivially be seen as an $O(n)$ algorithm.

3. Proofs

ZCNT is a distance metric [4] which means that it satisfies three properties:

1. **Identity:** $dist_{ZCNT}(u, u) = 0$
2. **Positivity:** if $u \neq v$ then $dist_{ZCNT}(u, v) > 0$
3. **Symmetry:** $dist_{ZCNT}(u, v) = dist_{ZCNT}(v, u)$
4. **Triangularity:** $dist_{ZCNT}(u, v) \leq dist_{ZCNT}(u, w) + dist_{ZCNT}(w, v)$

Below we will prove each of these properties mathematically.

3.1. Identity

$$\begin{aligned} dist_{ZCNT}(u, u) &= \frac{1}{2}||\Delta(u) - \Delta(u)||_1 \\ &= \frac{1}{2}||\mathbf{0}||_1 \\ &= \frac{1}{2}(0) \\ dist_{ZCNT}(u, u) &= 0 \end{aligned}$$

3.2. Positivity

The proof of positivity is based on the principle that the L1 norm of any nonzero vector will be greater than 0. Since Δ is a bijective function [4], if $u \neq v$, then $\Delta(u) \neq \Delta(v)$ so $\Delta(u) - \Delta(v) \neq \mathbf{0}$. Therefore we get the following implication: if $u \neq v$, then $0 < ||\Delta(u) - \Delta(v)||_1$. Additionally since L1 norms are always non-negative, if $0 < ||\Delta(u) - \Delta(v)||_1$, then $0 < 0.5 * ||\Delta(u) - \Delta(v)||_1$. From this we get a string of implications to show that:

if $u \neq v$, then $0 < ||\Delta(u) - \Delta(v)||_1 = dist_{ZCNT}(u, v)$ and $0 < dist_{ZCNT}(u, v)$, satisfying positivity.

3.3. Symmetry

$$\begin{aligned}
dist_{ZCNT}(u,v) &= \frac{1}{2}||\Delta(u) - \Delta(v)||_1 \\
&= \frac{1}{2}|| - (\Delta(u) - \Delta(v))||_1 \\
&= \frac{1}{2}||\Delta(v) - \Delta(u)||_1 \\
dist_{ZCNT}(u,v) &= dist_{ZCNT}(v,u)
\end{aligned}$$

3.4. Triangularity

Let $\mathbf{x} = (\Delta(u) - \Delta(w))$ and $\mathbf{y} = (\Delta(w) - \Delta(v))$. From this we get:

$$\begin{aligned}
||\mathbf{x}||_1 &= ||\Delta(u) - \Delta(w)||_1 \\
||\mathbf{y}||_1 &= ||\Delta(w) - \Delta(v)||_1 \\
||\mathbf{x} + \mathbf{y}||_1 &= ||\Delta(u) - \Delta(w) + \Delta(w) - \Delta(v)||_1 \\
&= ||\Delta(u) - \Delta(v)||_1
\end{aligned}$$

Since L1 is a norm, we know $||\mathbf{x} + \mathbf{y}|| \leq ||\mathbf{x}|| + ||\mathbf{y}||$. Therefore $||\Delta(u) - \Delta(v)||_1 \leq ||\Delta(u) - \Delta(w)||_1 + ||\Delta(w) - \Delta(v)||_1$. Substituting in the distance formula, we get $2 * dist_{ZCNT}(u,v) \leq 2[dist_{ZCNT}(u,w) + dist_{ZCNT}(w,v)]$ so $dist_{ZCNT}(u,v) \leq dist_{ZCNT}(u,w) + dist_{ZCNT}(w,v)$, satisfying the triangular inequality.

4. Methods

Finally I would like to implement both Lazac and another parsimony method that uses CNT. The paper lists improvements when solving both the small and large parsimony problems. Then I would like to compare the performance of ZCNT and CNT based models on simulated data so I have ground truth solutions to assess performance.

The points of interest in this paper are:

1. How well does ZCNT replicate CNT? How well does it replicate CN3?

2. How does ZCNT behave when the corresponding CNT does not exist?

3. How is

5. Results

6. Discussion

7. Conclusion

References

[1] Mohammed El-Kebir. Lecture 2, cs598meb computational cancer genomics. Oral Presentation, 2024. 1

[2] Mohammed El-Kebir, Benjamin J. Raphael, Ron Shamir, Roded Sharan, Simone Zaccaria, Meirav Zehavi, and Ron Zeira. Complexity and algorithms for copy-number evolution problems. *Algorithms for Molecular Biology*, 12(1):13, 2017. 2

[3] D. Hanahan and R.A. Weinberg. Hallmarks of cancer: The next generation. *Cell*, 144(646-674), 2011. 1

[4] Henri Schmidt, Palash Sashittal, and Benjamin J Raphael. A zero-agnostic model for copy number evolution in cancer. *bioRxiv*, pages 2023–04, 2023. 1, 2

[5] Roland F. Schwarz, Anne Trinh, Botond Sipos, James D. Brenton, Nick Goldman, and Florian Markowetz. Phylogenetic quantification of intra-tumour heterogeneity. *PLOS Computational Biology*, 10(4):1–11, 04 2014. 1

[6] Ron Shamir, Meirav Zehavi, and Ron Zeira. A Linear-Time Algorithm for the Copy Number Transformation Problem. In Roberto Grossi and Moshe Lewenstein, editors, *27th Annual Symposium on Combinatorial Pattern Matching (CPM 2016)*, volume 54 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 16:1–16:13, Dagstuhl, Germany, 2016. Schloss Dagstuhl – Leibniz-Zentrum für Informatik. 2