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. This vector is referred to as either a *copy number state* or *copy number profile*. 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 [6].

While useful and consistent with biological observations, there is no efficient algorithm using CNT to generate parsimonious cancer phylogenies due to its lack of symmetry. In *A zero-agnostic model for copy number evolution in cancer* [5], a new model ZCNT is proposed that provides a mathematical simplification on the CNT model, allowing for an efficient algorithm to generate cancer phylogenies.

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 the spontaneous creation of a locus that did not exist previously or having a negative number of copies of DNA. 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.

ZCNT supposedly closely mimics CNT distance but adds symmetry in an efficient manner. This paper will discuss the implications of using a symmetric property to mimic a non-symmetric property.

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 when the DNA stored within cancer cells is altered in some way inconsistent with healthy somatic behavior. These mutations can take multiple different forms. Mutations are random and unique to each tumor but lead to the development of common hallmarks. These hallmarks are referred to as the "Hallmarks of Cancer" [4] and are considered to be behaviors necessary for the cancer cells to grow and develop.

Since these hallmarks are acquired through evolution, studying the phylogenies of cancer cells can help inform how the cells acquired their abilities. Phylogenies are trees that trace the evolutionary history of a cell. In cancer phylogenics, trees are defined such that the nodes are "clones". Clones are defined as the genetic makeup of a population of cancer cells with common mutations. Studying cancer phylogenies can provide insight into tumor behaviors and open new avenues for potential treatments.

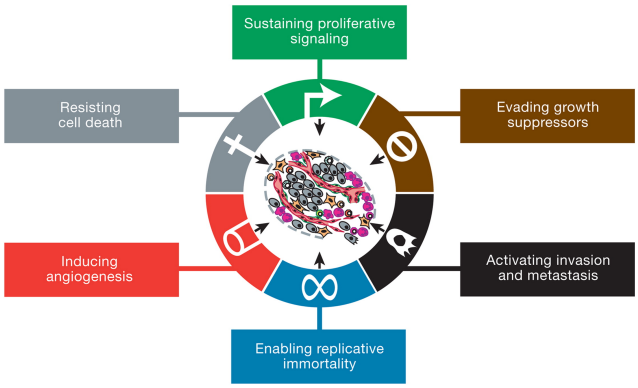

Figure 1: Hallmarks of Cancer [4]

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.

Researchers use genetic sequencing technologies to determine the state of SNVs and CNAs in a cell. The technology used determines what kind of analysis can be done with the data. If a technology has high coverage and low depth, it is considered optimal for determining SNVs in a cell. Meanwhile, if a technology has low coverage and high depth, it is considered optimal for determining the copy number state of a cell. There are currently no high-throughput technologies that provide sufficient data to do both SNV and CNA analysis. Therefore most approaches are geared towards one or the other. The CNT and ZCNT models that are discussed in this paper will focus on CNAs.

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 [8]. That means that in order to recon-

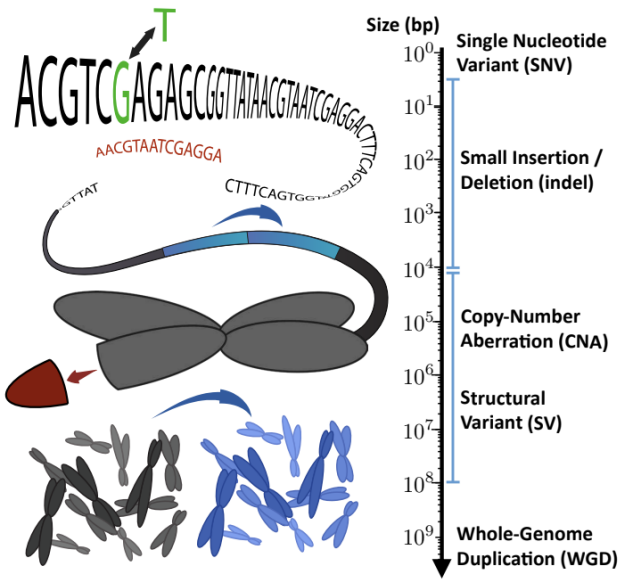

Figure 2: Different kinds of mutations affect genomes differently and their effects are sometimes of vastly different scales. This graphic relates different kinds of mutations to the scale of base pairs affected by a mutation of that type. [1]

struct cancer phylogenies, 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 the distance between two profiles to be the minimum number of events required to transform one profile into another ($\text{dist}(u, v) = \min_{T(u)=v} |T|$). Events in a transformation are commutative under ZCNT [5].

2.3.1 CNT

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

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

Given two copy number profiles, it is not guaranteed that there will be a transformation from one to the other under the CNT model. For instance, given $u = [0, 1]$ and $v = [1, 0]$, u would have to amplify a copy number from zero to one in order to transform it into v. In these cases, the distance is defined as $+\infty$. We will denote the CNT distance between two profiles u and v as $\mathbf{cnt}(u, v)$.

Due to the lack of symmetry in the CNT model, two common methods have been employed to create symmetric measures from CNT for various parsimony algorithms.

Both these algorithms are centered around getting the CNT distance in both directions. Mean Correction is when you take the average of the distances between two profiles. Median Distance just takes the lower of the two distances.

Method	Formula
Mean Correction	$\frac{\mathbf{cnt}(u,v)+\mathbf{cnt}(v,u)}{2}$
Median Distance	$\min\{\mathbf{cnt}(u, v), \mathbf{cnt}(v, u)\}$

Both of these definitions raise interesting questions about reachability. Mean Correction amplifies unreachability because if either $\mathbf{cnt}(u, v)$ or $\mathbf{cnt}(v, u)$ does not exist, the Mean Correction will yield $+\infty$. Meanwhile, Median Distance amplifies reachability because even if either $\mathbf{cnt}(u, v)$ or $\mathbf{cnt}(v, u)$ does not exist, Median Distance will yield a noninfinite number.

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 that minimizes $\mathbf{cnt}(w, u) + \mathbf{cnt}(w, v)$ [2].

While CN3 provides symmetry, it comes at the cost of computational time. CNT distances can be computed in $O(n)$ time [7] while the most optimal current algorithm for CN3 is $O(nB^7)$ where B is the largest copy number in the input profiles [2]. The benefit of using CN3 is that the value is symmetric and therefore works well with existing parsimony methods.

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 models 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 the ZCNT distance between two profiles u and v as $\mathbf{zcnt}(u, v)$. Unlike CNT distance, ZCNT distance always exists as a non-infinite number.

In *A zero-agnostic model for copy number evolution in cancer* [5], they propose the following closed-form formula for ZCNT distance:

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

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

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

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 proposed as a distance metric [5]. In order to be a distance metric, ZCNT must satisfy the four following properties:

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

Below we will prove each of these properties mathematically.

3.1. Identity

The identity property can be mathematically obtained without any additional lemmas.

Theorem 3.1 *Let u be some arbitrary copy-number profile under the ZCNT model.*
Then $\textbf{zcnt}(u, u) = 0$

Proof:

Statement	Explanation
$\textbf{zcnt}(u,u) = \frac{1}{2}\|\|\Delta(u) - \Delta(u)\|\|_1$	Equation 3
$= \frac{1}{2}\|\|\mathbf{0}\|\|_1$	Subtract
$= \frac{1}{2}(0)$	Simplify
$\textbf{zcnt}(u,u) = 0$	Simplify

3.2. Positivity

Theorem 3.2 *Let u and v be some arbitrary copy-number profiles under the ZCNT model.*
Then $u \neq v \implies 0 < \textbf{zcnt}(u, v)$.

Proof:
The proof of positivity is based on the principle that the L1 norm is a distance metric. Therefore by the positivity property, the L1 norm of any nonzero vector will be greater than 0. We can write this formally as:

Lemma 3.3 *Let u and v be some arbitrary vector in $\mathbb{R}^n$.*
Then $u \neq v \implies 0 < ||u - v||_1$.

Since Δ is a bijective function [5], we know that every input has a unique output. We also know that every output has a unique input. Therefore if $u \neq v$, then the delta mappings will be different. We can write this formally as:

Lemma 3.4 *Let u and v be some copy number profiles under the ZCNT model. Let $\Delta(\cdot)$ be the function to convert a copy number profile to its respective delta mapping.*
Then $u \neq v \iff \Delta(u) \neq \Delta(v)$

Using these properties, we can write the following proof for the positivity property of ZCNT

Statement	Explanation
$u \neq v \implies \Delta(u) \neq \Delta(v)$	Lemma 3.4
$\implies \Delta(u) - \Delta(v) \neq 0$	Subtraction
$\implies 0 < \|\|\Delta(u) - \Delta(v)\|\|_1$	Lemma 3.3
$\implies 0 < \frac{1}{2}\|\|\Delta(u) - \Delta(v)\|\|_1$	Multiply
$u \neq v \implies 0 < \textbf{zcnt}(u, v)$	Equation 3

3.3. Symmetry

Theorem 3.5 *Let u and v be some arbitrary copy-number profiles under the ZCNT model.*
Then $\textbf{zcnt}(u, v) = \textbf{zcnt}(v, u)$

Proof:

Corollary 3.5.1 *$\forall x \in \mathbb{R}$, $|x| = |-x|$ by absolute value definition*

Statement	Explanation
$\textbf{zcnt}(u,v) = \frac{1}{2}\|\|\Delta(u) - \Delta(v)\|\|_1$	Equation 3
$= \frac{1}{2}\|\| - (\Delta(u) - \Delta(v))\|\|_1$	Corollary 3.5.1
$= \frac{1}{2}\|\|\Delta(v) - \Delta(u)\|\|_1$	Multiply
$\textbf{zcnt}(u,v) = \textbf{zcnt}(v,u)$	Equation 3

3.4. Triangularity

Theorem 3.6 *Let u and v be some arbitrary copy number profiles under the ZCNT model.*
Then for all copy number profiles w,
$\textbf{zcnt}(u, v) \leq \textbf{zcnt}(u, w) + \textbf{zcnt}(w, v)$

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

$$\begin{aligned}
||\mathbf{x}||_1 &= ||\Delta(u) - \Delta(w)||_1 \\
&= 2 \times \textbf{zcnt}(u, w) \\
||\mathbf{y}||_1 &= ||\Delta(w) - \Delta(v)||_1 \\
&= 2 \times \textbf{zcnt}(w, v) \\
||\mathbf{x} + \mathbf{y}||_1 &= ||\Delta(u) - \Delta(w) + \Delta(w) - \Delta(v)||_1 \\
&= ||\Delta(u) - \Delta(v)||_1 \\
&= 2 \times \textbf{zcnt}(u, v) \qquad (5)
\end{aligned}$$

Since L1 is a distance metric we can write its triangularity property formally as follows:

Lemma 3.7 $\forall x, y \in \mathbb{R}^n, ||\mathbf{x}+\mathbf{y}||_1 \leq ||\mathbf{x}||_1 + ||\mathbf{y}||_1$

Using these two equations, we can formally prove the triangularity property of ZCNT as follows:

Statement	Explanation
$\|\|\mathbf{x}+\mathbf{y}\|\|_1 \leq \|\|\mathbf{x}\|\|_1 + \|\|\mathbf{y}\|\|_1$	Lemma 3.7
$2 \times \mathbf{zcnt}(u,v) \leq 2[\mathbf{zcnt}(u,w) + \mathbf{zcnt}(w,v)]$	Substitutions 5
$\mathbf{zcnt}(u,v) \leq \mathbf{zcnt}(u,w) + \mathbf{zcnt}(w,v)$	Divide

4. Methods

Since ZCNT is a metric that is meant to emulate CNT, the question of behavior across symmetry exists. CNT is not a symmetric property and doesn't always exist. If ZCNT typically approximates CNT, how does it behave when the corresponding CNT distance does not exist? Does it trend towards $+\infty$ when $\mathbf{cnt}(u,v) = \infty$ or is there another behavior?

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 often do reachability issues show up in simulated data?
4. How often do reachability issues show up in real data?
5. Does ZCNT more closely mimic Median Distance or Mean Correction?

4.1. Comparing ZCNT and CNT

In order to compare ZCNT and CNT distances, pairwise distances were generated for simulated data. The data was the same simulated data from the original ZCNT paper [5]. There were three kinds of distances generated for every simulation pair: ZCNT, CNT, and CN3.

The distances were filtered so we only considered the distances where CNT for that pairing was not $+\infty$. Since both ZCNT and CN3 always exist at a noninfinite value, this was the only filter. Then the relative error was calculated for each distance pairing (all combinations of ZCNT, CNT, and CN3).

4.2. Behavior of ZCNT when CNT is infinite

In order to determine ZCNT behavior when CNT is infinite, the same simulated distances in Section 4.1 were used to classify copy-number pairings into three classes:

Class 1: $\mathbf{cnt}(u,v)$ and $\mathbf{cnt}(v,u)$ both exist
Class 2: $\mathbf{cnt}(u,v)$ exists but $\mathbf{cnt}(v,u)$ does not
Class 3: Neither $\mathbf{cnt}(u,v)$ nor $\mathbf{cnt}(v,u)$ exist

Then an analysis was done based on the distribution of the corresponding ZCNT distances.

4.3. Reachability in Simulated Data

The same simulated data from the prior two sections was put through Lazac, a ZCNT Large Parsimony algorithm [5]. Then, using the calculated pairwise distances in CNT and ZCNT, two statistics were taken:

1. The percentage of edges in the overall tree represent biologically infeasible transformations.
2. The percentage of ancestor-descendant relationships that were biologically infeasible.

We define "biologically infeasible" transformations to be transformations wherein a copy-number goes from zero to a positive number. These can easily be determined by calculating the CNT distances. If the CNT distance between two nodes is ∞, the transformation is biologically infeasible.

For ancestor-descendant relationships between an ancestor u and a descendant v, the pairing is considered illegal if and only if an edge drawn straight between them would be biologically infeasible.

4.4. Reachability in Real Data

In order to determine how often reachability issues show up in real data, the same classification process described in Section 4.2 will be performed on real and simulated data. Bar graphs will be constructed to see if real data has a different class distribution than simulated data.

4.5. ZCNT Against Symmetric CNT Distances

In order to determine ZCNT's behavior against symmetric CNT measures like Median Distance or Mean Correction, ZCNT and CNT distances will be calculated for simulated data. Then the error between ZCNT and the different symmetric CNT measurements will be calculated and compared.

In order for figures to be legible, cases where the symmetric measures equal ∞ will be excluded. To make sure that the data is still represented, the number of omitted values will be included for each distance measure.

5. Results

5.1. Comparing ZCNT and CNT

The results for the methods described in Section 4.1 to compare the three distance algorithms ZCNT, CNT, and CN3 are shown in Figures 3 and 4 include five suites with a constant number of 200 cells and varying numbers of loci. Most other analyses also compare across varying numbers of cells and random seeds. This analysis does not do so due to the running time involved in calculating CN3 distances. A figure demonstrating this running time is included as Figure 5. This significantly higher runtime was expected due to CN3 being an $O(nB^7)$ algorithm while ZCNT and CNT are both linear time $O(n)$ algorithms.

Figures 3 and 4 show that ZCNT and CN3 have more in common than any other pairing. However, the relative error for the grand majority of the comparison classes falls below 10%. Interestingly, even the most different pairing (ZCNT and CNT) mostly had errors falling below 20%.

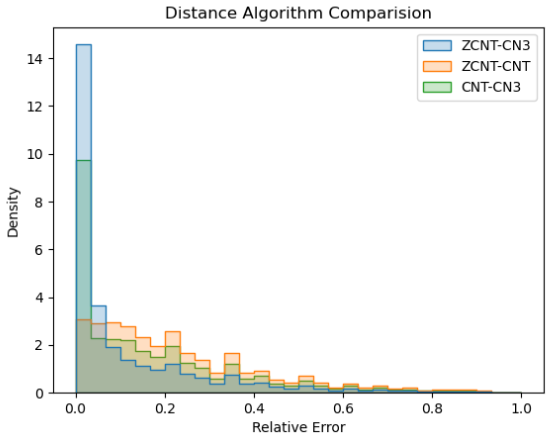

Figure 3: Relative error for the pairwise-distances of CNT, CN3, and ZCNT methods on simulated data as described in Section 4.1. This includes errors from multiple test suites where the number of cells was 200 and the loci was in {1000, 2000, 3000, 4000}. The graph was then normalized so that the area under the curve for each comparison class would equal 1. Pairings where the corresponding CNT distance was ∞ were excluded from the graph.

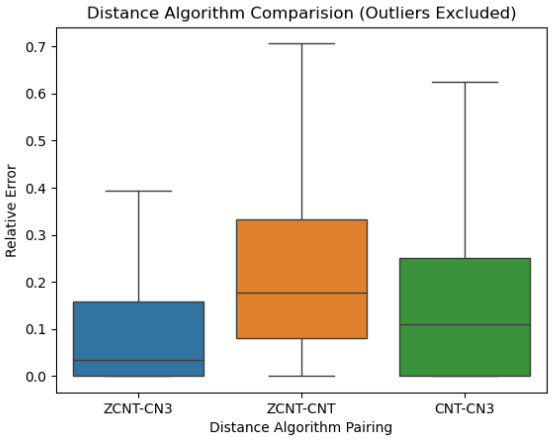

Figure 4: Relative error for the pairwise-distances of CNT, CN3, and ZCNT methods on simulated data as described in Section 4.1. Note that outliers are removed from this plot. This includes errors from multiple test suites where the number of cells was 200 and the loci was in {1000, 2000, 3000, 4000}. Pairings where the corresponding CNT distance was ∞ were excluded from the graph.

5.2. Behavior of ZCNT when CNT is infinite

Figures 6 and 7 show the results from the methodology described in Section 4.2. The three classes seem to all follow fairly normal distributions. Class 1 has the lowest mean while class 3 has the highest mean. This seems to signify to that reachability in CNT corresponds to lower ZCNT values. This is good because if trees minimize distance, they will also likely favor pairings in Class 1 over all other classes and favor pairings in Class 2 over pairings in Class 3. This behavior would align with biologically feasible transitions.

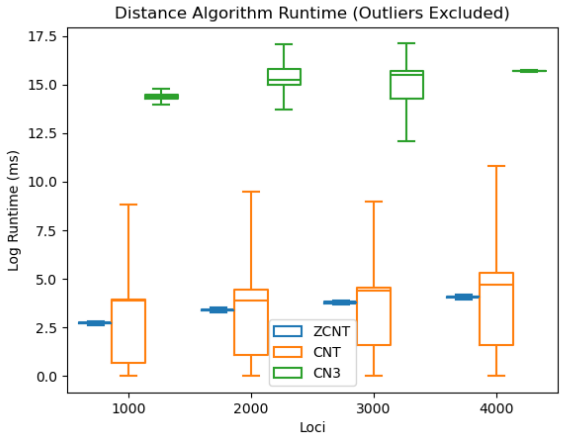

Figure 5: This shows the running times of running the different distance methods on different numbers of loci on a logarithmic scale. All trials used 200 cells. This was computed using simulated data. The number of loci was in {1000, 2000, 3000, 4000}.

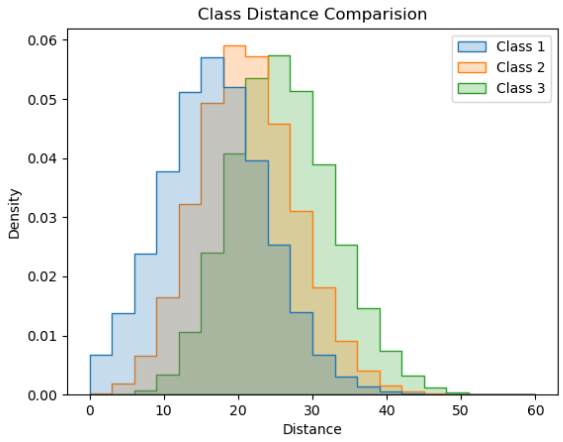

Figure 6: Absolute value of pairwise distances generated with ZCNT according to Section 4.2. Class 1 is composed of ZCNT distances $\mathbf{zcnt}(u, v)$ where $\mathbf{cnt}(u, v) \neq \infty$ and $\mathbf{cnt}(v, u) \neq \infty$. Class 2 is composed of ZCNT distances $\mathbf{zcnt}(u, v)$ where exactly one of $\mathbf{cnt}(u, v)$ and $\mathbf{cnt}(v, u)$ equals ∞. Class 3 is composed of ZCNT distances where $\mathbf{cnt}(u, v) = \mathbf{cnt}(v, u) = \infty$. The distances are from simulated data where the number of cells was in {200 300 400 500 600 1200} and the number of loci was in {1000, 2000, 3000, 4000} across seven random seeds. The data was then binned and normalized so the area under each curve is 1.

5.3. Reachability in Simulated Data

Figures 8 and 9 show the results from the methodology described in Section 4.3. These show that as the number of loci increases, biologically infeasible edges become less common. The number of cells involved in the phylogeny appears to decrease the spread of the data, but not shift it. It also shows that biologically infeasible edges are relatively rare, with 22 out of 168 trees (13.09%) having no illegal edges or ancestor-descendant relationships. The rate of illegal edges is very in-

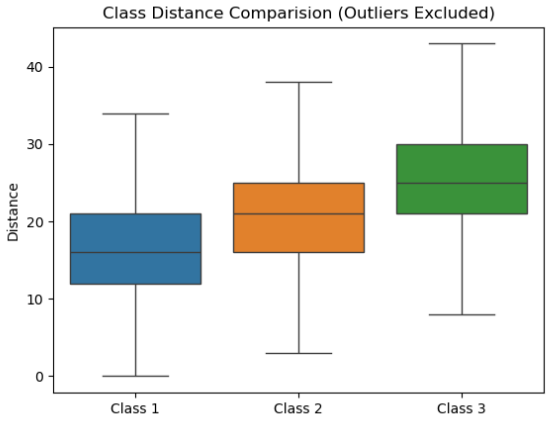

Figure 7: Absolute value of pairwise distances generated with ZCNT according to Section 4.2. Note that outliers are removed from this plot. Class 1 is composed of ZCNT distances $\mathbf{zcnt}(u, v)$ where $\mathbf{cnt}(u, v) \neq \infty$ and $\mathbf{cnt}(v, u) \neq \infty$. Class 2 is composed of ZCNT distances $\mathbf{zcnt}(u, v)$ where exactly one of $\mathbf{cnt}(u, v)$ and $\mathbf{cnt}(v, u)$ equals ∞. Class 3 is composed of ZCNT distances where $\mathbf{cnt}(u, v) = \mathbf{cnt}(v, u) = \infty$. The distances are from simulated data where the number of cells was in {200 300 400 500 600 1200} and the number of loci was in {1000, 2000, 3000, 4000} across seven random seeds.

frequent with is very small. Even the tree with the most illegal edges has at most 7.36% of illegal edges.

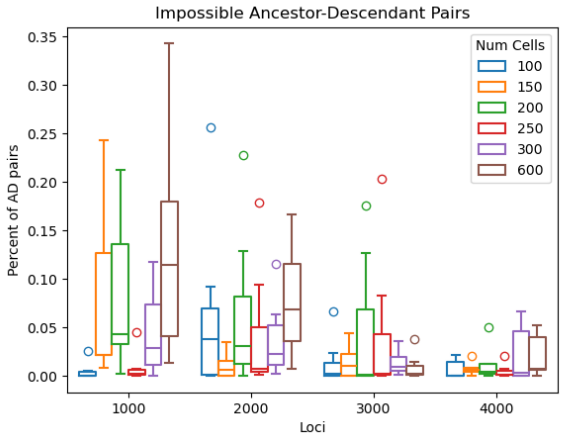

Figure 8: The percentage of overall ancestor-descendant pairs in Lazac-produced trees on simulated data where it would be biologically infeasible for the descendant to be produced from the ancestor. The number of cells was in {200 300 400 500 600 1200} and the number of loci was in {1000, 2000, 3000, 4000} across seven random seeds.

Figure 10 shows the cumulative rate of illegal edges or ancestor-descendant relationships. Illegal edges are the source of illegal ancestor-descendant relationships so the ancestor-descendant curve is just a horizontally stretched version of the illegal edges curve. The amount of stretch shows how many cells are descended from illegal edges.

The plot shows that more than half the trees have fewer than

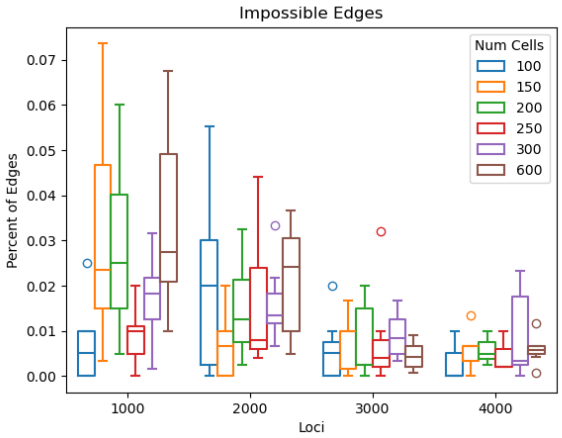

Figure 9: The percentage of overall edges in Lazac-produced trees on simulated data where it would be biologically infeasible for the child to be produced from the parent. The number of cells was in {200 300 400 500 600 1200} and the number of loci was in {1000, 2000, 3000, 4000} across seven random seeds.

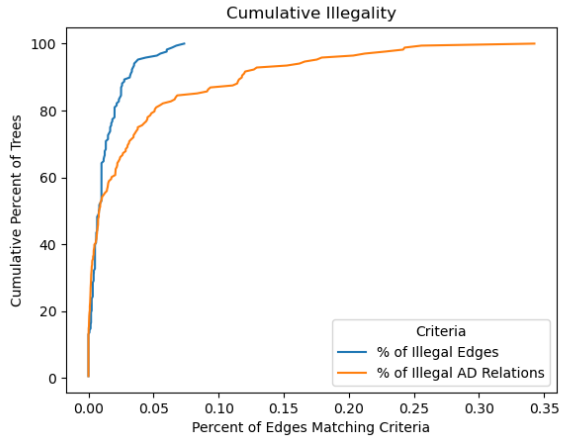

Figure 10: This shows a cumulative plot that shows what percentage of simulated trees contain biologically infeasible edges or ancestor-descendent relationships. The number of cells was in {200 300 400 500 600 1200} and the number of loci was in {1000, 2000, 3000, 4000} across seven random seeds.

2.5% illegal edges. The trees with the fewest illegal edges don't appear to have many descendants of those illegal relationships as shown by the curves matching for the first portion of the graph. This implies that for those trees, the ZCNT model predicts decently biologically feasible trees. However, for roughly 40% of the trees, an increasing proportion of nodes are descended from impossible transitions like zero-amplifications.

5.4. Reachability in Real Data

Figures 11 and 12 show the results from the methods described in Section 4.4. They show that the number of pairings falling into Classes 1 and 3 are about equal in real data but the simulated data shows a preference for Class 1.

For both kinds of data, the frequency of pairings falling into Class 2 is the highest but the peak is stronger in the real data.

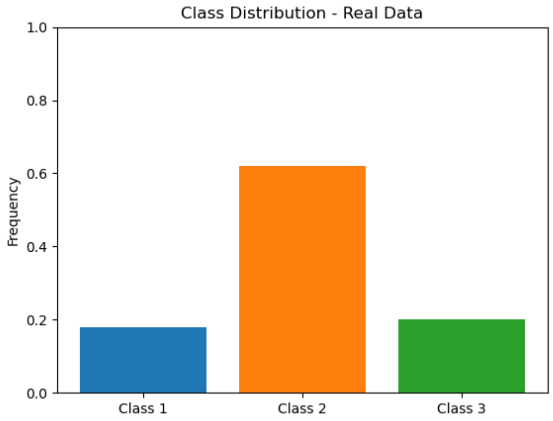

Figure 11: The frequency distribution of classes described in Section 4.2 using real data. The real data is from patient 8 of the study on metastatic prostate cancer [3]. The data was measured across 10 cells, a maximum of 22 chromosomes per cell, and a maximum of 38 loci per chromosome. Pairwise CNT distances were generated for each cell. Then each pairing was classified as follows. Class 1 is composed of pairings where $\mathbf{cnt}(u, v) \neq \infty$ and $\mathbf{cnt}(v, u) \neq \infty$. Class 2 is composed of pairings where exactly one of $\mathbf{cnt}(u, v)$ and $\mathbf{cnt}(v, u)$ equals ∞. Class 3 is composed of pairings where $\mathbf{cnt}(u, v) = \mathbf{cnt}(v, u) = \infty$.

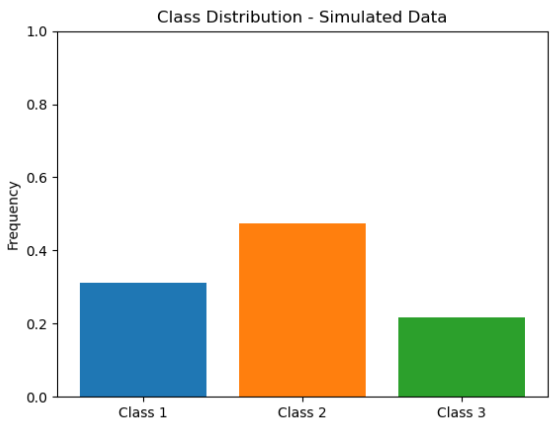

Figure 12: The frequency distribution of classes described in Section 4.2 using simulated data. The simulated data was generated such that the number of cells was in {200 300 400 500 600 1200} and the number of loci was in {1000, 2000, 3000, 4000} across seven random seeds. Pairwise CNT distances were generated for each suite and then classified as follows. Class 1 is composed of pairings where $\mathbf{cnt}(u, v) \neq \infty$ and $\mathbf{cnt}(v, u) \neq \infty$. Class 2 is composed of pairings where exactly one of $\mathbf{cnt}(u, v)$ and $\mathbf{cnt}(v, u)$ equals ∞. Class 3 is composed of pairings where $\mathbf{cnt}(u, v) = \mathbf{cnt}(v, u) = \infty$.

This could be due to several factors including the fact that there is significantly less data available for the real data compared to the simulated data.

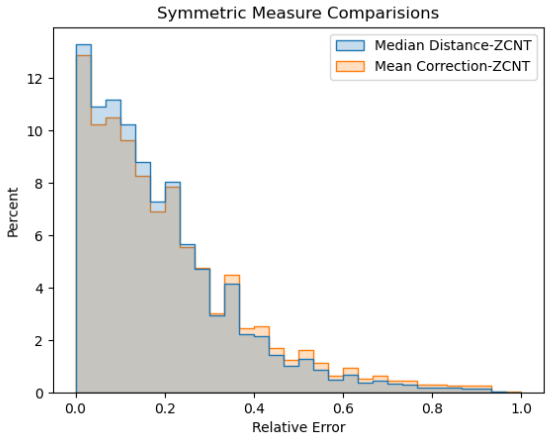

Figure 13: The relative error of ZCNT against CNT measures that correct for symmetry (Median Distance and Mean Correction) on simulated data as described in Section 4.5. The distances are from simulated data where the number of cells was in 200 300 400 500 600 1200 and the number of loci was in 1000, 2000, 3000, 4000 across seven random seeds. Pairwise ZCNT and CNT distances were generated for all cells in the data. Then Mean Correction and Median Distance were calculated for all CNT distances. Pairings where the relative distance would be ∞ due to Mean Correction or Median Distance being ∞ were excluded.

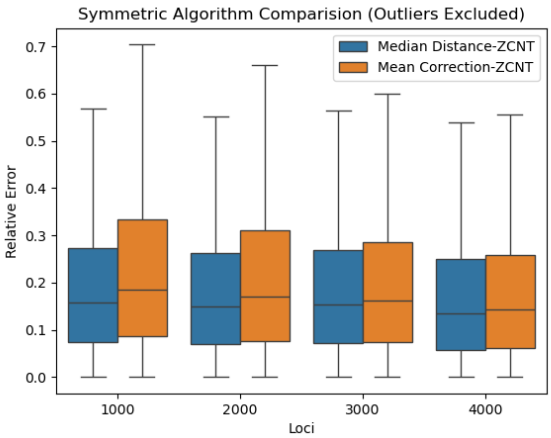

Figure 14: The relative error of ZCNT against CNT measures that correct for symmetry (Median Distance and Mean Correction) on simulated data as described in Section 4.5. Note that outliers are excluded from the graph. The distances are from simulated data where the number of cells was in 200 300 400 500 600 1200 and the number of loci was in 1000, 2000, 3000, 4000 across seven random seeds. Pairwise ZCNT and CNT distances were generated for all cells in the data. Then Mean Correction and Median Distance were calculated for all CNT distances. Pairings where the relative distance would be ∞ due to Mean Correction or Median Distance being ∞ were excluded.

5.5. ZCNT Against Symmetric CNT Distances

Figures 13 and 14 show the results of the methods described in Section 4.5. Figure 13 shows that ZCNT's relative error with

Loci	Mean Correction	Median Distance	Total
1000	76.57%	27.67%	52.12%
2000	76.06%	29.34%	52.70%
3000	51.22%	8.91%	30.07%
4000	44.51%	5.88%	25.19%
Total	62.09%	17.95%	40.02%

Figure 15: The percentage of pairwise distances excluded from plots in Figures 13 and 14 due to the value being ∞ rounded to the nearest hundredth of a percent.

Mean Correction and with Median Distance are nearly identical. Figure 14 shows that the number of loci in the sample does not strongly affect the relative error between ZCNT and central CNT measurements.

Figure 15 shows what percentage of pairs were excluded for each kind of distance measurement. Since ZCNT always exists as a noninfinite value, this is a representation of how often Mean Correction and Median Distance came up with infinite values. It seems that the more loci there are, the smaller the percentage of pairings result in infinite values regardless of the approach. This is interesting because the number of loci does not affect relative error. This means that the excluded values due to infinite Median Distance or Mean Correction values do not strongly affect the distribution of the distances.

The percentage of pairings resulting in infinite values is substantially higher for Mean Correction than it is for Median Distance. This is to be expected because for any distance from profile u to v, both $\mathbf{cnt}(u, v)$ and $\mathbf{cnt}(v, u)$ must be noninfinite for the Mean Correction to be noninfinite. Meanwhile for Median Distance, only one of them must be noninfinite for the resulting value to be noninfinite.

6. Discussion

Despite the basis of ZCNT being a biologically inaccurate phenomenon, the results from using this approach seem to produce biologically accurate results. Results from this paper show that biologically impossible phenomena like zero-amplification are improbable in ZCNT-based models.

This paper's research questions (See list 4) are answered by our analyses as follows:

1. The analysis presented in Section 5.1 shows that ZCNT closesly replicates both CN3 and CNT distance measurements. However, it appears to more closely mimic CN3 distance than CNT distance. This is perhaps due to the shared symmetric nature of ZCNT and CN3 algorithms.

2. The analysis presented in Section 5.2 shows that ZCNT trends lower when the corresponding CNT exists in both directions, marginally higher when unreachable in one direction, and marginally higher still if unreachable in both directions.

3. The analysis presented in Section 5.3 shows that the more data points available, the more biologically feasible the results from ZCNT Large Parsimony trees will be. More loci means that the number of illegal edges and ancestor-descendant relationships will decrease. This is perhaps because more loci increase the spread of the ZCNT distance, decreasing the average cost per edge and causing a stronger preference for easily reachable profiles. More analysis could be done in this avenue to determine if this is the case.

4. The analysis presented in Section 5.4 shows that the reachability distributions in real data are roughly the same as the reachability distributions in simulated data. The difference appears to be in the spread of the reachability with real data having a larger difference in frequency. More analysis could be done in this avenue to see if these differences are present when run on a larger collection of real data.

5. The analysis presented in Section 5.5 shows that ZCNT adheres to Median Distance and Mean Correction approximately the same amount regardless of how many values either measure deems to be infinite.

Code used to generate and analyze data can be found at https://github.com/sonalimerchia/CS598MEB-FinalProject

References

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

[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] Gunes Gundem, Peter Van Loo, Barbara Kremeyer, Ludmil B. Alexandrov, Jose M. C. Tubio, Elli Papaemmanuil, Daniel S. Brewer, Heini M. L. Kallio, Gunilla Högnäs, Matti Annala, Kati Kivinummi, Victoria Goody, Calli Latimer, Sarah O'Meara, Kevin J. Dawson, William Isaacs, Michael R. Emmert-Buck, Matti Nykter, Christopher Foster, Zsofia Kote-Jarai, Douglas Easton, Hayley C. Whitaker, David E. Neal, Colin S. Cooper, Rosalind A. Eeles, Tapio Visakorpi, Peter J. Campbell, Ultan McDermott, David C. Wedge, G. Steven Bova, and ICGC Prostate UK Group. The evolutionary history of lethal metastatic prostate cancer. *Nature*, 520(7547):353–357, 2015. 7

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

[5] 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, 3, 4

[6] 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

[7] 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

[8] I P Tomlinson, M R Novelli, and W F Bodmer. The mutation rate and cancer. *Proc Natl Acad Sci U S A*, 93(25):14800–14803, Dec. 1996. 1