International Journal of Advances in Engineering and Management (IJAEM)
Volume 6, Issue 11 Nov. 2024, pp: 145-153 www.ijaem.net ISSN: 2395-5252

Jukes-Cantor Correction for Phylogenetic Tree Reconstruction

Emunefe Friday Gabriel [‡], Ugbene Ifeanyichukwu Jeff[†]

[†] Department of Mathematics, Federal University of Petroleum Resources, Effurun.
[‡] Department of Mathematics, Federal University of Petroleum Resources, Effurun.

Date of Submission: 01-11-2024 Date of Acceptance: 10-11-2024

ABSTRACT: Phylogenetic tree reconstruction relies on accurate estimation of evolutionary distances between sequences. However, the observed Hamming distance between sequences can be misleading due to saturation, where multiple substitutions at the same site obscure the true evolutionary history. The Jukes-Cantor correction method addresses this by accounting for multiple substitutions, providing a more accurate representation of evolutionary distance. This study investigates the application of the Jukes-Cantor correction to the Hamming distance of genetic sequences in a case study, highlighting its impact on phylogenetic tree reconstruction. Our results demonstrate that the Jukes-Cantor correction significantly improves the accuracy of phylogenetic inference, particularly for sequences with substan tial evolutionary divergence. However, the model's reliance on simplifying assumptions, such as equal substitution rates and lack of base composition bias, limits its applicability to sequences with moderate levels of divergence. This study stands as a bedrock for further research into more complex models that can account for model violations and provide more accurate estimations of evolutionary distances for highly divergent sequences.

Keywords: Phylogenetic tree reconstruction, Jukes-Cantor correction, Hamming distance, saturation, evolutionary distance, model violations.
AMS Subject Classification 2010: 92D15 , 92B05, 62P10, 05C05, 68W05.

I. INTRODUCTION

Phylogenetic tree reconstruction is a critical aspect of evolutionary biology, providing insights into the evolutionary relationships among different species or genetic sequences. Among the various methods available for constructing phylogenetic trees, distance-based methods are widely used due to their sim plicity and computational efficiency. The Jukes-Cantor (JC) correction is one such method, which ac counts for multiple substitutions at a single site, thereby providing a more accurate estimation of evo lutionary distances. This study aims to explore the application of the Jukes-Cantor correction in recon structing phylogenetic trees, highlighting its significance and effectiveness in evolutionary studies.

Phylogenetic analysis has evolved significantly over the years, with various methods being developed to infer evolutionary relationships. One of the pioneering works in this field was the development of distance-based methods, which rely on the calculation of genetic distances between sequences [18]. These methods are favored for their computational efficiency and ease of implementation. The Jukes-Cantor model, introduced by Jukes and Cantor (1969) [14], is a fundamental approach in molecular evolution that assumes equal probability for all types of nucleotide substitutions. This model corrects for multiple hits at the same site, providing a more accurate distance estimate compared to simple p-distance methods. The Jukes-Cantor correction has been widely adopted in phylogenetic studies due to its robustness and simplicity [17].

Several studies have demonstrated the effectiveness of the Jukes-Cantor model in phylogenetic tree reconstruction. For instance, Tamura et al. (2004) [19] compared various distance correction methods and found that the Jukes-Cantor model consistently produced reliable phylogenetic trees, especially for closely related sequences. Similarly, Kumar et al. (2018) [16] highlighted the importance of using cor rected distance measures, including the Jukes-Cantor model, to avoid underestimation of evolutionary distances. Ane et al. (2007) [1] introduced Bayesian estimation techniques to assess concordance among gene trees, providing valuable insights into evolutionary relationships. Benson et al. (2008) [2] discussed the importance of Genbank in storing genetic information and its relevance to phylogenetic studies. Bor dewich et al. (2009) [3] explored the consistency of topological moves based on the balanced minimum evolution principle, shedding light on the inference of phylogenetic relationships. DeBry (1992) [4] in vestigated the consistency of phylogeny-inference methods under varying evolutionary rates, offering a comprehensive analysis of the challenges in evolutionary studies. Dowling et al. (2003) [5] compared a priori and a posteriori methods in studying host-parasite associations, emphasizing the significance of different approaches in evolutionary research. Edgar (2004) [6] developed the Muscle algorithm for multiple sequence alignment, enhancing the accuracy of genetic analyses. The works of Felsenstein (1978) [7], Ge et al. (1999) [8], and Harris (2019) [9] provided essential insights into phylogenetic analysis, taxonomy, and evolutionary relationships. These studies, along with others such as Herberts et al. (2022) [11], Henning (1966) [10] and Huelsenbeck et al. (1997) [12], have contributed to the understanding of evolutionary processes and the reconstruction of phylogenetic trees.

However, it is essential to acknowledge the limitations of the Jukes-Cantor model. While it provides a useful correction for multiple substitutions, it assumes equal base frequencies and substitution rates, which may not hold true for all datasets [21]. Advanced models such as the Kimura 2-parameter and the General Time Reversible (GTR) model have been developed to address these limitations by incorporating variable substitution rates and base frequencies [15, 20].

Despite these advancements, the simplicity and effectiveness of the Jukes-Cantor correction continue to make it a popular choice for phylogenetic analysis, particularly for preliminary studies and datasets with relatively uniform base compositions. This study aims to build on the existing literature by applying the Jukes-Cantor correction to reconstruct phylogenetic trees, evaluating its performance amongst other distance correction methods.

II. MATHEMATICAL FORMULATION

A phylogenetic tree is a graphical representation of the evolutionary relationships between a set of or ganisms or genes. It depicts the inferred evolutionary history of these entities, showing their common ancestors and the branching patterns that led to their diversification. A phylogenetic tree can be defined as a directed or undirected graph T = (V,E) where: V is the set of vertices, representing the taxa (or-ganisms or genes) being studied, and E is the set of edges, representing the evolutionary relationships between the taxa. A rooted tree has a designated root vertex representing the most recent common ances tor of all taxa in the tree. Edges are directed away from the root, indicating the direction of evolutionary descent. On the other hand, an unrooted tree does not have a designated root vertex. It only shows the relationships between taxa without specifying a common ancestor. Edges are undirected, representing evolutionary relationships without a defined direction of descent.

Let us consider two phylogenetic trees denoted as $T = (V,E)$ and $T^0 = (V^0,E^0)$. Given that T and T^0 possess specific properties and that isomorphisms of directed trees maintain indegrees and outdegrees, and preserve degrees for undirected trees, a function $\psi : T \rightarrow T^0$can only be an isomorphism of the phylogenetic trees X and X^0if ψ forms a bijection $\psi : X \rightarrow X^0$ on the sets of leaf nodes. Thus, it is necessary that $|X| = |X^0|$. In the context of biology, an isomorphism of phylogenetic trees, represented by $\varphi : T \rightarrow T^0$, implies that the restriction $\varphi : X \rightarrow X^0$ of $\varphi :V \rightarrow V^0$acts as an identity map, indicating that $X = X^0$and $\varphi(v) = v$ for all $v \in X$. This concept of isomorphism elucidates how different representations of phylogenetic trees can convey the same evolutionary relationships among the leaf nodes.

Consider the unrooted binary phylogenetic tree $T_1 = ((A,B),(C,D))$ for $X = \{A,B,C,D\}$. In this tree, the common ancestor of the pairs {A,B} and {C,D} is denoted as v, while the ancestor of the remaining pairs is denoted as u. Another unrooted binary phylogenetic tree $T_2 = ((A,C),(B,D))$is defined, featuring the ancestor s for the pair {A,C} and the ancestor t for the pair {B,D}. An isomorphism between T_1 and T_2 as phylogenetic trees can be established through the mapping $\varphi : T_1 \rightarrow T_2$ with assignments such as $\varphi(A) = C$, $\varphi(B) = D$, $\varphi(u) = s$, $\varphi(v) = t$, $\varphi(C) = A$, and $\varphi(D) = B$. Notably, the focus here lies on the structural relationships, disregarding edge lengths.

While phylogenetic trees inherently possess labeled leaf nodes, the addition of labels to the edges can enhance phylogenetic tree reconstruction. Interpreting the vertices V of a phylogenetic tree T = (V,E) as species, edge labels can convey information about evolutionary changes between species. In graph theory, labeling the edges E of T is termed as edge-weighting, defined by a function $\omega : E \rightarrow R$ assigning a real value to each edge $e \in E$. Edge-weightings are commonly nonnegative, but flexibility in allowing broader edge-weightings can benefit phylogenetic tree reconstruction algorithms. The concept of edge-weighting in phylogenetics aligns with an evolutionary distance map, crucial for determining evolutionary distances through models explaining sequence changes. The study of evolutionary distances is a fundamental aspect of biological and biomathematical research, with extensive literature available for further exploration.

In the course of our analysis, we will generate trees $T \in T_n$ and associated weightings ω using distance-based reconstruction methods. The collection of ordered pairs comprising unrooted binary phy logenetic X-trees T and positive edge weightings ω is denoted as $T_n = \{(T,\omega)|T = (V,E) \in T_n, \omega : E \rightarrow R^+\}$. Extending T_n to encompass edge weightings with zero or negative values from certain reconstruc tion techniques could offer further insights and advancements in phylogenetic tree analysis.

Phylogenetic trees often incorporate branch lengths, which represent the amount of evolutionary change that has occurred along each branch. These lengths can be measured in various units, such as: Genetic distance, which is the number of nucleotide substitutions or amino acid changes between two taxa. This is denoted by T(u, v), the path (sequence of edges) connecting vertices u and v in the tree. The distance (branch length) between vertices u and v denoted by d(u, v) , is measured along the path T(u, v).

2.1 Distance Methods

Distance methodologies utilize a collection of pairwise distances between sequences in a specified re duced multiple alignment to reconstruct trees, which can be either rooted or unrooted depending on the methodology employed. It is assumed that these distances are provided without detailing their specific derivation process. However, we will later delve into a common approach for generating distances, or more precisely, alternative values for distances that we term as ”pseudodistances.” Initially, we present a formal definition. Consider M as a set, and let $d : M \times M \rightarrow R$ be a function. We define d as a distance function on M if it satisfies the following conditions:

1. $d(u, v) > 0$ for all $u, v \in M$, where u 6= v,

2. $d(u,u) = 0$ for all $u \in M$,

3. $d(u, v) = d(v,u)$ for all $u, v \in M$,

4. The triangle inequality is upheld: $d(u, v) \leq d(u,w) + d(w, v)$ for all $u, v,w \in M$.

A metric space is defined as a set equipped with a distance function adhering to the specified conditions and phylogenetic trees are likely to .

The value d(u, v) representing any pair of $u, v \in M$ is denoted as the distance between u and v when d operates as a distance function on M. By introducing a distance function on M, we have the ability to transform any set M into a metric space. This transformation involves defining d(u, v) = 1 for all $u, v \in M$ where u 6= v, and setting d(u,u) = 0 for all $u \in M$. However, this particular distance function offers limited informational value. Our focus will be on the specific scenario of distance functions applied to a finite assortment $M = \{x_1,..., x_N\}$ of genetic sequences intended for phylogenetic tree construction. Let us suppose that a distance function d is established on M, with d encapsulating insights into the extent of divergence among the sequences within M. This implies that d holds biological significance. For example, if sequences x_i and x_j have diverged further from their common ancestor compared to x_k and x_l, then $d(x_i, x_j) > d(x_k, x_l)$. For ease of notation, we will denote $d(x_i, x_j)$ as $d_{i\ j}$. Utilizing the symmetric distance matrix $M_d = (d_{i\ j})$ will be beneficial in representing the information encoded by d.

The distance $d_T(x_i, x_j) = d_{Ti\ j}$ in tree T represents the length of the shortest path from $x_i$to x_j. By establishing an unrooted tree T connecting the genetic sequences, a tree-induced distance function d_T is generated on M. It is shown that, under broad assumptions, d_T qualifies as a distance function on M. The primary objective of distance methodologies in phylogenetic analysis is to identify all trees T where the distance function d_T closely approximates d. Such trees are deemed optimal in the realm of distance methodologies. Consequently, the essence of distance methodologies lies in determining branch lengths and unrooted trees collectively (while also addressing a technique that constructs rooted trees).

It logically ensues that if a tree T exists that produces the distance function d, then $d_T = d$ ($d_{Ti\ j} = d_{i\ j}$ for all i, j), establishing d as an additive distance function on M. For the case of N = 2, the response to this inquiry is unequivocally affirmative. Let us now consider the scenario where N = 3. In this case, the three sought-after positive values u, v,w are such that

$$u+v=d_{12},$$
$$u+w=d_{13}$$
$$v+w=d_{23}.$$

(1)

The solution to equations (1) is

$$u=\frac{1}{2}(d_{12}+d_{13}-d_{23}),$$
$$v=\frac{1}{2}(d_{12}+d_{23}-d_{13}),$$
$$w=\frac{1}{2}(d_{13}+d_{23}-d_{12}).$$

(2)

We notice that due to the triangle inequality, the quantities on the right side of equation (2) are non negative. While they do not necessarily have to be positive, as the inequality is not strict, some of them could indeed be equal to 0. For this reason, we opt to allow for the presence of zero branch lengths, assuming all branch lengths to be non-negative values moving forward, rather than strictly positive. In biological contexts, branches with zero length are considered ”very short” branches. As the definition of additivity remains consistent with the previously provided definition, equation (2) illustrates that any distance function is additive on M in this broader sense when N = 3.

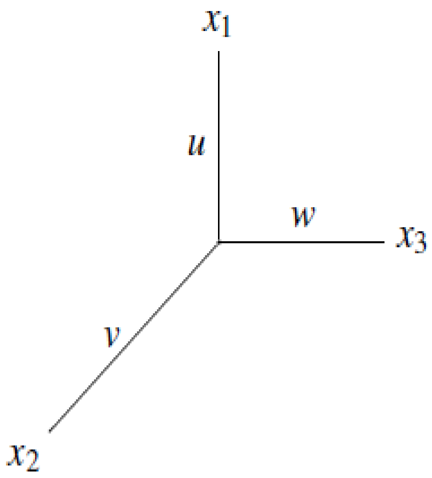

Figure 1: Phylogenetic tree of 3 unknown genetic sequences

At times, we set this requirement independently because the distance function d_T may not meet condition (1) of the definition of a distance function if certain branch lengths in a tree T are zero. It is important to note that with the allowance of zero branch lengths, phylogenetic trees can exhibit any branching pattern at internal nodes, as opposed to solely following the bifurcating pattern discussed earlier. As observed, there exists only one tree that generates the specified distance function for N = 2,3. In the realm of additive distance functions, the uniqueness of such a tree is a commonly acknowledged fact.

2.2 Jukes-Cantor correction to the Hamming distance

The number of positions in which two sequences, denoted as x and y, exhibit differences is referred to as the Hamming distance, denoted as $d_H(x, y)$. Consider the scenario where we are presented with two sequences, x and y, composed of elements from the set {A,G,C,T}.

x = G A T T C A T T C
y = G C C A T A T T C.

Hence the Hamming distance $d_H(x, y)$ between x and y is 4.
The Jukes-Cantor correction d_{JC} to the Hamming distance is defined as

$$d_{JC}(x,y)=-\frac{3}{4}\log\left(1-\frac{4}{3}f\right),$$

(3)

Assuming f denotes the frequency of unique sites that differentiate between two sequences, consider the above scenario where we have sequences x and y each of length 9, with a Hamming distance of 4, denoted as $d_H(x, y) = 4$. Consequently, we find f $=\frac{4}{9}$ = 0.4444. An elementary yet rudimentary ap proach to quantifying sequence dissimilarity is through the application of the Hamming distance. This method overlooks possibilities such as character modifications over time and potential reversals in spe cific instances. Additionally, it fails to consider established biological principles, like the non-uniform likelihood of a DNA character transitioning into another, influenced by the specific DNA bases and their arrangement in the sequence. The term ”evolutionary models” pertains to particular additional as sumptions and techniques utilized to determine the evolutionary distances between two given leaves, represented by aligned sequences (DNA, RNA, proteins, etc.), denoted as x and y. These assumptions and techniques are employed to address various challenges. Notably, s_x and s_y are contingent on the selection of evolutionary models.

On a collection M, suppose d acts as a

distance function, and let N ≥ 4. In this case, d is deemed additive if and only if the following condition is satisfied: for any set of four distinct numbers 1 ≤ i, j, k,l ≤ N, the two sums that are equal and greater than or equal to the third sum are $d_{i\ j}$ +d_{kl}, d_{ik} +d_{jl}, and d_{il} + d_{jk}. Subsequently, a traceback procedure is employed to construct the tree. This method involves keeping track of which pair of genetic sequences from the preceding step resulted in a specific genetic sequence at the current step [13, 18]. Further elaboration on the algorithm will now be provided. Define, for each i = 1,...,N,

$$r_i = \frac{1}{N-2}\sum_{k=1}^{N} d_{ik}. \tag{4}$$

Further, for all i, j = 1,...,N, i < j, set

$$D_{ij} = d_{ij} - (r_i + r_j). \tag{5}$$

manner:

$$d_{N+1m} = \frac{1}{2}|d_{im} + d_{jm} - d_{ij}| \tag{7}$$

We are now able to iterate the previously outlined procedure with the updated set of N − 1 genetic sequences M^0 = {x_m, x_{N+1},m 6= i, j}. Following these iterations, a single unrooted tree topology emerges, continuing until only three genetic sequences remain, at which stage the associated branch lengths are computed utilizing formulas (2). Subsequently, a traceback operation is employed to construct the tree.

III. RESULT

In this section, we will be applying the methods discussed in the previous section to analyze case studies and obtain meaningful results. By so doing, it allows us to reconstruct the evolutionary relationships between the observed entities in a rather intriguing manner, minimizing the number of evolutionary events required. By applying this method, we aim to gain comprehensive insights into the underlying structure and patterns present in phylogenetic structures.

Now lets consider six (6) DNA sequences the set X = {A,G,T,C}, as entailed below;

x_1 = A T C G A T C G A T C G A T
x_2 = A T C G A T C G A T C G A A

We can represent $D_{i\ j}$ in an upper-triangular matrix D = ($D_{i\ j}$) for convenience. Let’s select a pair where $D_{i\ j}$ is the minimum for 1 ≤ i, j ≤ N (not necessarily unique). The genetic sequences x_i, x_j will then be merged into a single group, replacing them with an genetic sequence x_{N+1} comprising a single element. The new genetic sequence x_{N+1} is situated at specific distances from x_i and x_j, serving as an internal node in the forthcoming tree:

$$\begin{aligned} d_{N+1i} &= \frac{1}{2}|d_{ij} + r_i - r_j|, \\ d_{N+1j} &= \frac{1}{2}|d_{ij} + r_j - r_i| \end{aligned} \tag{6}$$

We shall proceed to establish the distances between x_{N+1} and any x_m where m 6= i, j in the subsequent

x_3 = A T C G A T C G A T C A A A
x_4 = A T C G A T C G A A C A A A
x_5 = T A C G T A C G T A C G T C
x_6 = C A T G T A C G T A C G T A

Now we get the distance matrix by computing the hamming distance between these sequences, this gives;

$$\begin{array}{ccccccc} M_d & x_1 & x_2 & x_3 & x_4 & x_5 & x_6 \\ x_1 & 0 & 1 & 2 & 3 & 8 & 9 \\ x_2 & 1 & 0 & 1 & 2 & 8 & 8 \\ x_3 & 2 & 1 & 0 & 1 & 9 & 9 \\ x_4 & 3 & 2 & 1 & 0 & 8 & 8 \\ x_5 & 8 & 8 & 9 & 8 & 0 & 3 \\ x_6 & 9 & 8 & 9 & 8 & 3 & 0 \end{array}$$

The Jukes-Cantor correction M_{dJC} distance matrix can be gotten as a correction to the hamming distance matrix using the equation (3), to give;

M_{dxc}	x_1	x_2	x_3	x_4	x_5	x_6
x_1	0.0000	0.0751	0.1585	0.2524	1.0763	1.4594
x_2	0.0751	0.0000	0.0751	0.1585	1.0763	1.0763
x_3	0.1585	0.0751	0.0000	0.0751	1.4594	1.4594
x_4	0.2524	0.1585	0.0751	0.0000	1.0763	1.0763
x_5	1.0763	1.0763	1.4594	1.0763	0.0000	0.2524
x_6	1.4594	1.0763	1.4594	1.0763	0.2524	0.0000

To ensure that D adheres to the criteria of a legitimate distance function, it is crucial to validate the four-point condition before initiating the neighbor-joining algorithm. However, in this instance, we will proceed with the neighbor-joining algorithm without conducting this validation process. A tree T will be constructed, and the derived function D_T will be compared against D. This comparison will demonstrate that D_T = D, affirming that D effectively fulfills the four-point condition.

$$r_1=\frac{3.0217}{4}=0.7554, r_2=\frac{2.4613}{4}=0.6153, r_3=\frac{3.2275}{4}=0.8069,$$
$$r_4=\frac{2.6386}{4}=0.6597, r_5=\frac{4.9407}{4}=1.2352, r_6=\frac{5.3238}{4}=1.3310$$

This gives the following matrix D':

D'	x_1	x_2	x_3	x_4	x_5	x_6
x_1		−1.2957	−1.4038	−1.1627	−0.9143	−0.6270
x_2			−1.3471	−1.1165	−0.7742	−0.8699
x_3				−1.3914	−0.5826	−0.6784
x_4					−0.8185	−0.9143
x_5						−2.3137

In the matrix provided, the smallest value is D_{13} = −1.4038. We will now introduce a fresh sequence denoted as x_7, which will take the position of the pair x_1, x_3. The placement of x_7 will be at a distance

$$d_{71}=\frac{1}{2}|d_{31}+r_1-r_3|=\frac{|0.1585+0.7554-0.8069|}{2}=\frac{0.1070}{2}=0.0535$$

from x_1 and at the distance

$$d_{73}=\frac{1}{2}|d_{31}+r_3-r_1|=\frac{|0.1585+0.8069-0.7554|}{2}=\frac{0.2100}{2}=0.1050$$

from x_3, as shown in Fig. 2.

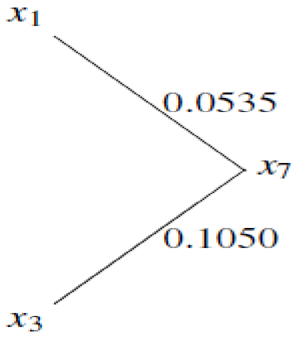

Figure 2:

We will now compute distances between x_7 and each of x_2, x_4, x_5, x_6. We have

$$d_{72}=\frac{1}{2}|d_{12}+d_{32}-d_{13}|=\frac{0.0083}{2}=0.0042,$$
$$d_{74}=\frac{1}{2}|d_{14}+d_{34}-d_{13}|=\frac{0.1711}{2}=0.0856$$
$$d_{75}=\frac{1}{2}|d_{15}+d_{35}-d_{13}|=\frac{2.3772}{2}=1.1886$$
$$d_{76}=\frac{1}{2}|d_{16}+d_{36}-d_{13}|=\frac{2.7603}{2}=1.3802$$

which gives the following distance matrix for the genetic sequences x_2, x_4, x_5, x_6, x_7:

M_{dxc}	x_2	x_4	x_5	x_6	x_7
x_2	0.0000	0.1585	1.0763	1.0763	0.0042
x_4	0.1585	0.0000	1.0763	1.0763	0.0856
x_5	1.0763	1.0763	0.0000	0.2524	1.1886
x_6	1.0763	1.0763	0.2524	0.0000	1.3802
x_7	0.0042	0.0856	1.1886	1.3802	0.0000

For this new distance matrix, we will repeat the process again and obtain

$$r_2=\frac{2.3069}{3}=0.7690, r_4=\frac{2.3967}{3}=0.7989,$$

$$r_5 = \frac{3.5936}{3} = 1.1979, r_6 = \frac{3.7852}{3} = 1.2617, r_7 = \frac{2.6502}{3} = 0.8834$$

which gives the following matrix:

$$\begin{array}{cccccc}
D & x_2 & x_4 & x_5 & x_6 & x_7 \\
x_2 & & -1.4094 & -0.8905 & -0.9544 & -1.6566 \\
x_4 & & & -0.9205 & -0.9843 & -1.5967 \\
x_5 & & & & -2.2072 & -0.8927 \\
x_6 & & & & & -0.7649
\end{array}$$

We now introduce a new genetic sequence, x_8, that will replace the pair x_5, x_6 (note that D_{56} is minimal in the above matrix). We place x_8 at a distance 0.0943 from x_5 and x_6 at a distance 0.1581 from x_8, as shown in Figure 3.

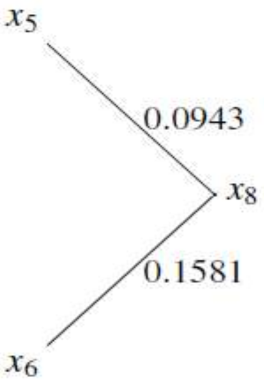

Figure 3:

The distance matrix for the sequences, x_2, x_4, x_7, x_8 is:

$$\begin{array}{ccccc}
M_{d_{JC}} & x_2 & x_4 & x_7 & x_8 \\
x_2 & 0.0000 & 0.1585 & 0.0042 & 0.9501 \\
x_4 & 0.1585 & 0.0000 & 0.0856 & 0.9501 \\
x_7 & 0.0042 & 0.0856 & 0.0000 & 1.1582 \\
x_8 & 0.9501 & 0.9501 & 1.1582 & 0.0000
\end{array}$$

On the next step of the algorithm, we obtain r_2 = 0.5522, r_4 = 0.5971, r_7 = 0.6198, r_8 = 1.5292, and:

$$\begin{array}{ccccc}
D & x_2 & x_4 & x_7 & x_8 \\
x_2 & & -0.9908 & -1.1762 & -1.1313 \\
x_4 & & & -1.1313 & -1.1762 \\
x_7 & & & & -0.9908
\end{array}$$

At this point, we can group together either x_2 and x_7, or x_4 and x_8, since both D_{27} and D_{48} are minimal in the above matrix (the resulting tree will not depend on our choice).

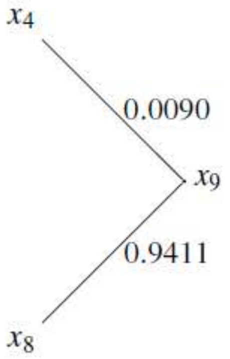

Figure 4:

We group together x_4 and x_8, that is, we introduce a new sequence x_9, place it at a distance 0.0090 from x_4 and at a distance 0.9411 from x_8, as shown in Figure 4, and calculate distances from x_9 to x_2 and x_7, which gives the following distance matrix for the three sequences:

$$\begin{array}{cccc}
M_{d_{JC}} & x_2 & x_7 & x_9 \\
x_2 & 0.0000 & 0.0042 & 0.0793 \\
x_7 & 0.0042 & 0.0000 & 0.1469 \\
x_9 & 0.0793 & 0.1469 & 0.0000
\end{array}$$

Going on now to determine the minimal pair from the above, r_2 = 0.0835, r_7 = 0.1511 and r_9 = 0.2262, so that

$$\begin{array}{cccc}
D & x_2 & x_7 & x_9 \\
x_2 & & -0.2304 & -0.2304 \\
x_7 & & & -0.2304
\end{array}$$

Jukes-Cantor Correction ... 11 From the above the we could pick any as the minimal pair, suppose we pick D_{27}

$$d_{102} = \frac{1}{2}|d_{27} + r_2 - r_7| = 0.0317,$$
$$d_{107} = \frac{1}{2}|d_{72} + r_7 - r_2| = 0.0359. \qquad (8)$$

We shall proceed to establish the distances between x_{10} and any x_9, in subsequent manner:

$$d_{109} = \frac{1}{2}\left|d_{29}+d_{79}-d_{27}\right| = 0.1110 \tag{9}$$

Then we introduce a new sequence x_{10}, such that

$$\begin{array}{ccc} M_{d_{JC}} & x_9 & x_{10} \\ x_9 & 0.0000 & 0.1110 \\ x_{10} & 0.1110 & 0.0000 \end{array}$$

It follows that the above distance function is generated by the tree shown in Figure 5.

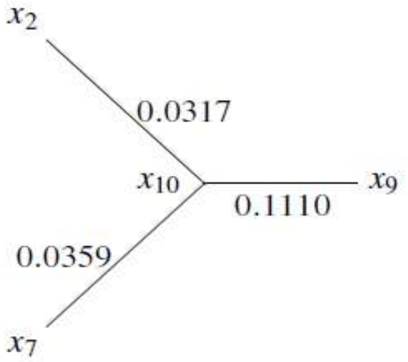

Figure 5:

Next, merging the trees from Figs. 2 – 5, we obtain the tree T shown in Fig. 6.

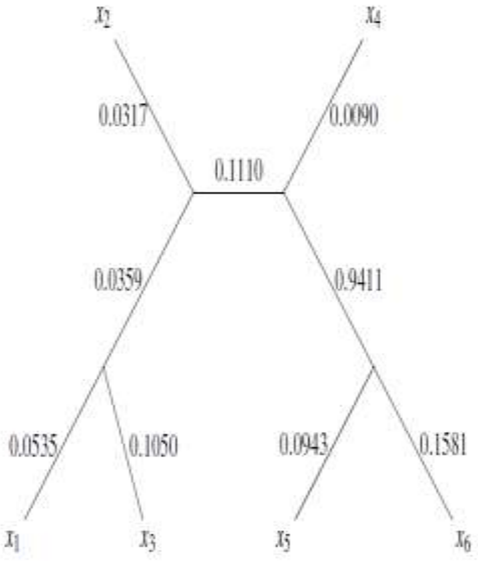

Figure 6:

It is easy to verify that T generates d and therefore the distance function d indeed satisfies the four point condition. The application of the Jukes-Cantor correction method to the Hamming distance of genetic sequences in our case study has yielded valuable insights into the accuracy and limitations of this approach in phylogenetic tree reconstruction. The Jukes-Cantor correction effectively addresses the issue of saturation in genetic distances, which occurs as evolutionary time increases and the observed Hamming distance plateaus, failing to reflect the true evolutionary distance. By accounting for multiple substitutions at the same site, the correction provides a more accurate estimation of the true evolutionary distance, simplifying phylogenetic analysis and allowing us to compare sequences that have undergone different levels of evolutionary change. This leads to more reliable tree topologies and branch lengths, as demonstrated by our case study, where the Jukes-Cantor correction significantly improved the accuracy of phylogenetic tree reconstruction, particularly when dealing with sequences that have experienced substantial evolutionary divergence.

However, the Jukes-Cantor model relies on several simplifying assumptions, including equal rates of substitution for all nucleotides and a lack of base composition bias. These assumptions may not always hold true in real-world scenarios, potentially leading to inaccuracies in distance estimation. Additionally, the Jukes-Cantor correction is most effective for sequences with relatively low levels of divergence. As the number of substitutions increases, the model's accuracy can decline, and more complex models, such as the Kimura 2-parameter model, may be necessary for highly divergent sequences. Furthermore, the accuracy of the correction is sensitive to violations of the model's assumptions. For example, if there is a significant base composition bias, the correction may underestimate the true evolutionary distance.

Future research could investigate the performance of other phylogenetic models, such as the Kimura 2-parameter model or the general time-reversible (GTR) model, in correcting for saturation and improv ing phylogenetic tree reconstruction. Developing methods to account for model violations, such as base composition bias, would further enhance the accuracy of phylogenetic analysis. Additionally, combining the Jukes-Cantor correction with other phylogenetic methods, such as Bayesian inference or maximum likelihood analysis, could lead to more robust and informative phylogenetic inferences.

IV. CONCLUSION

The Jukes-Cantor correction method is a valuable tool for addressing saturation in genetic distances and improving the accuracy of phylogenetic tree reconstruction. While it relies on simplifying assumptions and may have limitations, it provides a robust and widely applicable method for analyzing moderate levels of sequence divergence. By understanding its strengths and limitations, researchers can utilize the Jukes-Cantor correction effectively to gain insights into evolutionary relationships and reconstruct phylogenetic trees with greater confidence.

REFERENCES

[1]. Ane, C., Larget, B., Baum, D., Smith, S., & Rokas, A. (2007). Bayesian estimation of concordance among gene trees. Mol Biol Evol, 24:412–426.

[2]. Benson, D., Karsch-Mizrachi, I., Lipman, D., Ostell, J., & Wheeler, D. (2008). Genbank. Nucleic Acids Res, 36:D25–D30.

[3]. Bordewich, M., Gascuel, O., Huber, K., & Moulton, V. (2009). Consistency of topological moves based on the balanced minimum evolution principle of phylogenetic inference. IEEE/ACM Trans Comput Biol Bioinform, 6:110–117.

[4]. DeBry, R. (1992). The consistency of several phylogeny-inference methods under varying evolu tionary rates. Mol Biol Evol, 9:537–551.

[5]. Dowling, A., Veller, M., Hoberg, E., & Brooks, D. (2003). A priori and a posteriori methods in comparative evolutionary studies of host-parasite associations. Cladistics, 19:240–253.

[6]. Edgar, R. (2004). Muscle: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res, 32:1792–1797.

[7]. Felsenstein, J. (1978). Cases in which parsimony or compatibility methods will be positively mis leading. Syst Zool, 22:240–249.

[8]. Ge, S., Sang, T., Lu, B., & Hong, D. (1999). Phylogeny of rice genomes with emphasis on origins of allotetraploid species. PNAS, 96:14400–14405.

[9]. Harris, K. (2019). Taxonomy & phylogeny. Biology LibreTexts. Retrieved 19 April 2023. [10] Henning, W. (1966). Phylogenetic systematics. Univ. of Illinois Press, Urbana, IL.

[10]. Herberts, C., Annala, M., Sipola, J., Ng, S. W. S., Chen, X. E., Nurminen, A., Korhonen, O. V., Munzur, A. D., Beja, K., Sch¨onlau, E., Bernales, C. Q., Ritch, E., Bacon, J. V. W., Lack, N. A., & Nykter, M. (2022). Deep whole-genome ctdna chronology of treatment-resistant prostate cancer. Nature, 608(7921):199–208.

[11]. Huelsenbeck, J., Rannala, B., & Yang, Z. (1997). Statistical tests of host-parasite cospeciation. Evolu- tion, 51:410–419.

[12]. Isaev, A., & Deem, M. (2005). Introduction to mathematical methods in bioinformatics. Physics Today, 58(10), 83-83. https://doi.org/10.1063/1.2138428

[13]. Jukes, T. H., & Cantor, C. R. (1969). Evolution of protein molecules. In H. N. Munro (Ed.), Mam malian Protein Metabolism (pp. 21-132). Academic Press.

[14]. Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. Journal of Molecular Evolution, 16(2), 111-120.

[15]. Kumar, S., Stecher, G., Li, M., Knyaz, C.,& Tamura, K. (2018). MEGA X: Molecular Evolutionary Genetics Analysis across computing platforms. Molecular Biology and Evolution, 35(6), 1547- 1549.

[16]. Nei, M., & Kumar, S. (2000). Molecular Evolution and Phylogenetics. Oxford University Press.

[17]. Saitou, N., & Nei, M. (1987). The neighbor-joining method: A new method for reconstructing phylogenetic trees. Molecular Biology and Evolution, 4(4), 406-425.

[18]. Tamura, K., Dudley, J., Nei, M., & Kumar, S. (2007). MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. Molecular biology and evolution, 24(8), 1596–1599.

[19]. Tavare, S. (1986). Some probabilistic and statistical problems in the analysis of DNA sequences. ´ Lectures on Mathematics in the Life Sciences, 17, 57-86.

[20]. Yang, Z. (1994). Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods. Journal of Molecular Evolution, 39(3), 306-314.