Substitution Models in Phylogenetic Reconstruction from Molecular Data: Theoretical Principles, Implementation and Selection Methodologies, Limits of Results Interpretation, and Recent Advances

Author: Richard Murdoch Montgomery

Affiliation: Scottish Science Society

Email: editor@scottishsciencesocietyperiodic.uk

Date: July 2025

Abstract

Substitution models constitute the mathematical foundation of modern phylogenetic inference, providing the probabilistic framework necessary for reconstructing evolutionary relationships from molecular sequence data. These models describe the stochastic processes governing nucleotide or amino acid changes over evolutionary time through continuous-time Markov chains, enabling maximum likelihood and Bayesian approaches to phylogenetic reconstruction. This comprehensive review examines the theoretical principles underlying substitution models, from the foundational Jukes-Cantor model to sophisticated general time-reversible frameworks, whilst addressing critical aspects of model selection, parameter estimation, and computational implementation. We analyse the mathematical formulations of key models including JC69, K80, F81, HKY85, and GTR, presenting their rate matrices, transition probabilities, and equilibrium conditions in formal notation suitable for publication. Through computational illustrations and graphical analyses, we demonstrate the behaviour of these models under varying parameter conditions and evolutionary scenarios. The article critically evaluates model selection

methodologies, including information-theoretic criteria and likelihood ratio tests, whilst discussing the inherent limitations and assumptions that constrain the interpretation of phylogenetic results. Recent advances in mixture models, partition-specific substitution processes, and machine learning approaches are examined in the context of improving phylogenetic accuracy and biological realism. Our analysis reveals that whilst substitution models have achieved remarkable sophistication, fundamental challenges remain in capturing the full complexity of molecular evolution, particularly regarding rate heterogeneity, non-stationarity, and epistatic interactions. The findings emphasise the importance of careful model selection and the recognition of model limitations in phylogenetic studies, providing guidance for practitioners in choosing appropriate methodological frameworks for their evolutionary analyses.

Keywords: substitution models, phylogenetic reconstruction, molecular evolution, maximum likelihood, Bayesian inference, model selection, DNA evolution, protein evolution, Markov chains, evolutionary genomics

---

1. Introduction

The reconstruction of phylogenetic relationships from molecular sequence data represents one of the most fundamental challenges in evolutionary biology, requiring sophisticated mathematical frameworks to infer the historical processes that have shaped the diversity of life on Earth. At the heart of this endeavour lie substitution models, which provide the probabilistic foundation for understanding how genetic sequences evolve over time through the accumulation of mutations, substitutions, and other molecular changes (Felsenstein, 1981). These models serve as the essential bridge between observable sequence differences and the underlying evolutionary processes, enabling researchers to quantify evolutionary distances, estimate divergence times, and reconstruct the branching patterns of phylogenetic trees with statistical rigour.

The development of substitution models has its origins in the pioneering work of Jukes and Cantor (1969), who introduced the first mathematical framework for correcting observed sequence differences to account for multiple substitutions at the same site. This seminal contribution recognised that the simple counting of differences between sequences systematically underestimates the true evolutionary distance due to the phenomenon of

homoplasy, where multiple changes at a single position can obscure the actual number of substitutional events that have occurred. The Jukes-Cantor model, whilst making simplifying assumptions about equal base frequencies and uniform substitution rates, established the fundamental principle that evolutionary inference requires probabilistic models that can account for the stochastic nature of molecular evolution.

The subsequent decades have witnessed an extraordinary expansion in the sophistication and biological realism of substitution models, driven by advances in molecular biology, computational power, and statistical methodology. The recognition that different types of nucleotide substitutions occur at different rates led to the development of models that distinguish between transitions and transversions, most notably the Kimura two-parameter model (Kimura, 1980). This advancement acknowledged the well-established observation that transitions between chemically similar bases occur more frequently than transversions between dissimilar bases, reflecting the underlying biochemical constraints on DNA replication and repair processes.

Further refinements incorporated the recognition that nucleotide frequencies are rarely equal in natural sequences, leading to models such as the Felsenstein F81 model that allow for unequal base compositions whilst maintaining other simplifying assumptions (Felsenstein, 1981). The integration of these insights culminated in the Hasegawa-Kishino-Yano model, which combines both transition-transversion bias and unequal base frequencies, providing a more realistic representation of the evolutionary process (Hasegawa et al., 1985). The ultimate generalisation of these approaches is embodied in the General Time Reversible model, which allows for the maximum number of free parameters whilst maintaining the crucial assumption of time reversibility that enables phylogenetic inference from contemporary sequences (Tavaré, 1986).

The mathematical foundation of substitution models rests upon the theory of continuous-time Markov chains, which provides the formal framework for describing the probabilistic evolution of discrete character states over continuous time intervals. In this formulation, the evolutionary process is characterised by an instantaneous rate matrix that specifies the rates of change between different character states, typically the four DNA bases or twenty amino acids. The fundamental assumption of the Markov property ensures that the probability of future changes depends only on the current state and not on the historical path by which that state was reached, a simplification that makes phylogenetic inference computationally tractable whilst capturing the essential features of molecular evolution.

The transition from instantaneous rates to finite-time transition probabilities is accomplished through matrix exponentiation, a mathematical operation that transforms the rate matrix into a probability matrix describing the likelihood of observing particular character states after a specified evolutionary time. This transformation is central to all likelihood-based phylogenetic methods, as it enables the calculation of the probability of observing particular sequence patterns given a hypothetical phylogenetic tree and set of model parameters. The computational challenges associated with matrix exponentiation have driven significant advances in numerical methods and algorithmic optimisation, making large-scale phylogenetic analyses feasible for contemporary datasets.

The application of substitution models extends far beyond simple distance estimation to encompass the full spectrum of phylogenetic inference methods. Maximum likelihood approaches utilise substitution models to calculate the probability of observing the data given a particular tree topology and set of parameters, enabling the identification of the most probable phylogenetic hypothesis through optimisation procedures (Yang, 2006). Bayesian methods incorporate substitution models within a probabilistic framework that allows for the quantification of uncertainty in phylogenetic estimates through the exploration of posterior probability distributions over tree space and parameter space (Huelsenbeck & Ronquist, 2001). Distance-based methods rely on substitution models to convert observed sequence differences into evolutionary distances that can be used in clustering algorithms such as neighbour-joining or UPGMA.

The selection of appropriate substitution models has emerged as a critical component of phylogenetic analysis, with significant implications for the accuracy and reliability of evolutionary inferences. The proliferation of available models has necessitated the development of systematic approaches to model selection, including information-theoretic criteria such as the Akaike Information Criterion and Bayesian Information Criterion, which balance model fit against model complexity to identify optimal parameterisations (Posada & Buckley, 2004). Likelihood ratio tests provide an alternative framework for comparing nested models through formal statistical hypothesis testing, whilst cross-validation approaches assess model performance through predictive accuracy on independent data subsets.

Despite the remarkable sophistication achieved by contemporary substitution models, fundamental limitations remain that constrain the accuracy and scope of phylogenetic inference. The assumption of rate constancy across sites and lineages, whilst

computationally convenient, fails to capture the extensive heterogeneity observed in real evolutionary processes. Among-site rate variation, arising from differences in functional constraints, structural requirements, and mutational mechanisms, requires sophisticated mixture models or gamma-distributed rate categories to achieve adequate representation (Yang, 1994). Similarly, the assumption of stationarity, which requires that the evolutionary process remains constant over time, is frequently violated in real datasets due to changes in selective pressures, population dynamics, and environmental conditions.

The challenge of model adequacy assessment has become increasingly important as phylogenetic datasets have grown in size and complexity. Traditional approaches to model validation, such as examination of residual patterns or simulation studies, provide limited insight into the specific ways in which models fail to capture biological reality. Recent advances in posterior predictive assessment and cross-validation methodologies offer more sophisticated approaches to model evaluation, enabling researchers to identify specific aspects of the data that are poorly explained by particular model formulations (Bollback, 2002).

The computational demands of modern phylogenetic analysis have driven significant innovations in algorithmic design and implementation. The calculation of likelihood functions for large datasets requires efficient algorithms for matrix exponentiation, numerical optimisation, and tree space exploration. Parallel computing architectures and graphics processing units have enabled analyses of unprecedented scale, whilst approximate methods such as composite likelihood approaches provide computational shortcuts for extremely large datasets at the cost of some statistical rigour (Stamatakis, 2014).

Recent developments in the field have focused on incorporating greater biological realism into substitution models whilst maintaining computational tractability. Codon-based models explicitly account for the genetic code and selection pressures on protein sequences, enabling more accurate inference of evolutionary processes in protein-coding regions (Goldman & Yang, 1994). Mixture models allow different sites or lineages to evolve under different substitution processes, capturing heterogeneity that is ignored by simpler models. Partition models enable different regions of the genome to evolve under distinct substitution processes, reflecting the diverse functional constraints operating on different genomic elements.

The integration of machine learning approaches with traditional substitution modelling represents an emerging frontier in phylogenetic methodology. Neural networks and other machine learning algorithms offer the potential to learn complex evolutionary patterns directly from data without requiring explicit specification of model structure, whilst maintaining the interpretability and statistical foundation that characterise traditional approaches (Suvorov et al., 2020). These hybrid methodologies may provide pathways to overcome some of the fundamental limitations of current substitution models whilst preserving the theoretical framework that enables rigorous statistical inference.

The epistemological challenges inherent in phylogenetic reconstruction reflect broader issues in contemporary multidisciplinary research, where traditional disciplinary boundaries may constrain our understanding of complex biological phenomena. As Montgomery (2025) argues in his analysis of ontological and epistemological frameworks in academic research, the integration of artificial intelligence and computational approaches in scientific inquiry requires careful consideration of both transformative potential and inherent limitations. The development of substitution models exemplifies this tension between mathematical sophistication and biological complexity, highlighting the need for frameworks that can accommodate uncertainty whilst providing actionable insights for evolutionary biology.

The practical application of substitution models in phylogenetic studies requires careful consideration of the trade-offs between model complexity, computational feasibility, and biological realism. Simple models may be adequate for certain applications whilst providing computational advantages and interpretability, whereas complex models may be necessary for accurate inference in challenging datasets despite their increased computational demands and parameter estimation difficulties. The choice of model must be guided by the specific objectives of the analysis, the characteristics of the dataset, and the available computational resources.

This comprehensive review aims to provide a thorough examination of substitution models in phylogenetic reconstruction, encompassing their theoretical foundations, mathematical formulations, implementation strategies, and practical applications. We present detailed mathematical derivations of key models, computational illustrations of their behaviour under different parameter regimes, and critical assessments of their strengths and limitations. The analysis includes extensive discussion of model selection methodologies, recent advances in model development, and future directions for the field. Through this

comprehensive treatment, we seek to provide both theoretical insights and practical guidance for researchers engaged in phylogenetic analysis, contributing to the continued advancement of evolutionary biology through improved methodological understanding and application.

2. Methodology

2.1 Theoretical Foundations

2.1.1 Continuous-Time Markov Chain Framework

The mathematical foundation of substitution models rests upon the theory of continuous-time Markov chains, which provides a rigorous probabilistic framework for describing the evolution of discrete character states over continuous time intervals. Consider a DNA sequence position that can exist in one of four possible states corresponding to the nucleotides A, G, C, and T, denoted as the state space $S = \{1, 2, 3, 4\}$ where the numerical indices correspond to the alphabetical ordering of bases. The evolutionary process is modelled as a continuous-time Markov chain $\{X(t) : t \geqslant 0\}$ where $X(t) \in S$ represents the nucleotide state at time t.

The fundamental assumption of the Markov property requires that the conditional probability distribution of future states depends only on the current state and not on the historical sequence of states that led to the current configuration. Formally, this is expressed as:

$$P(X(t+s) = j \mid X(t) = i, X(u) = x_u, 0 \leqslant u < t) = P(X(t+s) = j \mid X(t) = i)$$

for all $i, j \in S$, $t, s \geqslant 0$, and any sequence of past states $\{x_u\}$. This assumption enables the complete specification of the evolutionary process through the instantaneous rate matrix and eliminates the need to track the complete evolutionary history of each sequence position.

The instantaneous rate matrix $\mathbf{Q} = (q_{ij})$ defines the rates of change between different nucleotide states, where q_{ij} for $i \neq j$ represents the instantaneous rate of substitution from state i to state j. The diagonal elements are constrained by the requirement that each row sums to zero:

$$q_{ii} = -\sum_{j \neq i} q_{ij}$$

This constraint ensures that the total rate of change from any given state equals the sum of rates to all other states, maintaining probability conservation throughout the evolutionary process. The rate matrix **Q** completely characterises the substitution process and serves as the fundamental parameter of all substitution models.

2.1.2 Equilibrium Distribution and Stationarity

A crucial concept in substitution model theory is the equilibrium distribution $\boldsymbol{\pi} = (\pi_1, \pi_2, \pi_3, \pi_4)^T$, which represents the long-term stationary frequencies of the four nucleotides. The equilibrium distribution satisfies the fundamental equation:

$$\boldsymbol{\pi}^T \mathbf{Q} = \mathbf{0}^T$$

where $\mathbf{0}$ is the zero vector. This equation expresses the condition that at equilibrium, the rate of change into each state exactly balances the rate of change out of that state, resulting in stable long-term frequencies. The equilibrium distribution is unique for irreducible rate matrices and represents the limiting distribution of nucleotide frequencies as evolutionary time approaches infinity.

The assumption of stationarity requires that the substitution process remains constant over time, implying that the rate matrix **Q** and equilibrium distribution $\boldsymbol{\pi}$ do not change during the evolutionary period under consideration. Whilst this assumption is clearly violated over very long evolutionary timescales due to changes in selective pressures and environmental conditions, it provides a reasonable approximation for many phylogenetic analyses and enables the mathematical tractability necessary for practical implementation.

2.1.3 Time-Reversibility and Detailed Balance

Many substitution models incorporate the assumption of time-reversibility, which requires that the evolutionary process appears identical when viewed forwards or backwards in time. This assumption is not merely a mathematical convenience but reflects the fundamental principle that phylogenetic inference from contemporary sequences requires models that do not distinguish between ancestral and descendant states. Time-reversibility is formally expressed through the detailed balance condition:

$$\pi_i q_{ij} = \pi_j q_{ji}$$

for all i, j ∈ S. This equation states that the equilibrium flow from state i to state j exactly equals the equilibrium flow from state j to state i, ensuring that the process exhibits no net directional bias when viewed over long time periods.

The detailed balance condition has profound implications for the structure of substitution models, as it constrains the number of free parameters in the rate matrix. For a four-state DNA model, the detailed balance condition reduces the number of independent rate parameters from twelve to six, corresponding to the six possible pairs of nucleotides. This reduction in dimensionality significantly simplifies parameter estimation and model comparison procedures.

2.1.4 Transition Probability Matrix

The transition probability matrix $\mathbf{P}(t) = (P_{ij}(t))$ describes the probability of observing nucleotide j at time t given that nucleotide i was present at time 0. The relationship between the instantaneous rate matrix and the transition probability matrix is governed by the Kolmogorov forward equation:

$$d\mathbf{P}(t)/dt = \mathbf{P}(t)\mathbf{Q}$$

with the initial condition $\mathbf{P}(0) = \mathbf{I}$, where $\mathbf{I}$ is the identity matrix. The solution to this differential equation is given by the matrix exponential:

$$\mathbf{P}(t) = \exp(\mathbf{Q}t) = \Sigma_{n=0}^{\infty} ((\mathbf{Q}t)^n/n!)$$

The matrix exponential provides the fundamental link between the instantaneous rates of substitution and the finite-time transition probabilities that are directly observable in phylogenetic data. The computation of matrix exponentials represents one of the primary computational challenges in likelihood-based phylogenetic inference, requiring sophisticated numerical algorithms for efficient and accurate evaluation.

2.1.5 Eigenvalue Decomposition and Computational Methods

For rate matrices that can be diagonalised, the matrix exponential can be computed efficiently through eigenvalue decomposition. If $\mathbf{Q} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^{-1}$ where $\mathbf{\Lambda} = \text{diag}(\lambda_1, \lambda_2, \lambda_3, \lambda_4)$ contains the eigenvalues and $\mathbf{U}$ contains the corresponding eigenvectors, then:

$$\mathbf{P}(t) = \mathbf{U}\exp(\mathbf{\Lambda}t)\mathbf{U}^{-1} = \mathbf{U}\text{diag}(e^{\lambda_1 t}, e^{\lambda_2 t}, e^{\lambda_3 t}, e^{\lambda_4 t})\mathbf{U}^{-1}$$

This decomposition reduces the computation of the matrix exponential to the evaluation of scalar exponentials, providing significant computational advantages for repeated calculations with different values of t. The eigenvalues of the rate matrix have important biological interpretations, with the largest eigenvalue (which is always zero for properly normalised rate matrices) corresponding to the equilibrium distribution, and the remaining eigenvalues determining the rates of convergence to equilibrium.

2.2 Specific Substitution Models

2.2.1 Jukes-Cantor Model (JC69)

The Jukes-Cantor model represents the simplest substitution model, making the assumptions of equal nucleotide frequencies and equal substitution rates between all pairs of nucleotides (Jukes & Cantor, 1969). Under these assumptions, the equilibrium distribution is uniform: $\boldsymbol{\pi} = (0.25, 0.25, 0.25, 0.25)^T$, and the rate matrix takes the form:

$\mathbf{Q}_{JC} = \mu [[-3/4, 1/4, 1/4, 1/4], [1/4, -3/4, 1/4, 1/4], [1/4, 1/4, -3/4, 1/4], [1/4, 1/4, 1/4, -3/4]]$

where μ represents the overall substitution rate parameter. The transition probability matrix for the JC69 model has a closed-form analytical solution:

$P_{ij}(t) = \{1/4 + 3/4 \times \exp(-4\mu t/3) \text{ if } i = j; 1/4 - 1/4 \times \exp(-4\mu t/3) \text{ if } i \neq j\}$

The JC69 distance correction formula provides a method for estimating evolutionary distances from observed sequence differences. If p represents the proportion of sites that differ between two sequences, the JC69 distance estimate is:

$d_{JC} = -3/4 \times \ln(1 - 4p/3)$

This formula corrects for multiple substitutions at the same site, which become increasingly important as evolutionary distances increase.

2.2.2 Kimura Two-Parameter Model (K80)

The Kimura two-parameter model extends the JC69 framework by distinguishing between transitions (substitutions between purines A↔G or between pyrimidines C↔T) and transversions (substitutions between purines and pyrimidines) (Kimura, 1980). This distinction reflects the well-established observation that transitions occur more frequently than transversions due to biochemical constraints on DNA replication and repair processes.

The K80 model maintains the assumption of equal nucleotide frequencies but introduces a transition-transversion rate ratio parameter κ. The rate matrix is:

$\mathbf{Q}\_K80 = \mu [[-(1+\kappa)/4, \kappa/4, 1/4, 1/4], [\kappa/4, -(1+\kappa)/4, 1/4, 1/4], [1/4, 1/4, -(1+\kappa)/4, \kappa/4], [1/4, 1/4, \kappa/4, -(1+\kappa)/4]]$

The K80 distance correction requires separate estimation of the proportions of transitional (P) and transversional (Q) differences:

$d\_K80 = -1/2 \times \ln[(1-2P-Q) \times \sqrt{(1-2Q)}]$

where P and Q represent the observed proportions of transitions and transversions, respectively.

2.2.3 Felsenstein Model (F81)

The Felsenstein F81 model relaxes the assumption of equal nucleotide frequencies whilst maintaining equal rates for all substitution types (Felsenstein, 1981). This model recognises that natural DNA sequences often exhibit significant compositional bias, with important implications for evolutionary distance estimation and phylogenetic inference.

The F81 rate matrix incorporates unequal equilibrium frequencies $\boldsymbol{\pi} = (\pi\_A, \pi\_G, \pi\_C, \pi\_T)^T$:

$\mathbf{Q}\_F81 = \mu [[-(\pi\_G + \pi\_C + \pi\_T), \pi\_G, \pi\_C, \pi\_T], [\pi\_A, -(\pi\_A + \pi\_C + \pi\_T), \pi\_C, \pi\_T], [\pi\_A, \pi\_G, -(\pi\_A + \pi\_G + \pi\_T), \pi\_T], [\pi\_A, \pi\_G, \pi\_C, -(\pi\_A + \pi\_G + \pi\_C)]]$

The transition probabilities for the F81 model are:

$P\_ij(t) = \pi\_j + (\delta\_ij - \pi\_j) \times \exp(-\mu t)$

where δ_ij is the Kronecker delta function.

2.2.4 Hasegawa-Kishino-Yano Model (HKY85)

The HKY85 model combines the transition-transversion distinction of the K80 model with the unequal base frequencies of the F81 model, providing a more realistic representation of molecular evolution (Hasegawa et al., 1985). This model has become one of the most widely used substitution models due to its balance between biological realism and computational tractability.

The HKY85 rate matrix incorporates both κ and unequal base frequencies:

$\mathbf{Q}_{HKY} = \mu [[*, \kappa\pi_G, \pi_C, \pi_T], [\kappa\pi_A, *, \pi_C, \pi_T], [\pi_A, \pi_G, *, \kappa\pi_T], [\pi_A, \pi_G, \kappa\pi_C, *]]$

where the diagonal elements are determined by the row-sum constraint. The HKY85 model requires estimation of five parameters: four base frequencies (with the constraint $\Sigma \pi_i = 1$) and the transition-transversion ratio κ.

2.2.5 General Time-Reversible Model (GTR)

The General Time-Reversible model represents the most parameter-rich time-reversible substitution model for DNA sequences (Tavaré, 1986). The GTR model allows for six independent substitution rate parameters corresponding to the six possible pairs of nucleotides, whilst maintaining the detailed balance condition that ensures time-reversibility.

The GTR rate matrix can be written as:

$\mathbf{Q}_{GTR} = \mu [[*, a\pi_G, b\pi_C, c\pi_T], [a\pi_A, *, d\pi_C, e\pi_T], [b\pi_A, d\pi_G, *, f\pi_T], [c\pi_A, e\pi_G, f\pi_C, *]]$

where a, b, c, d, e, f are the six exchangeability parameters that determine the relative rates of different substitution types. The GTR model encompasses all simpler time-reversible models as special cases through appropriate parameter constraints.

2.3 Model Selection and Parameter Estimation

2.3.1 Maximum Likelihood Estimation

Parameter estimation in substitution models is typically accomplished through maximum likelihood methods, which seek to identify the parameter values that maximise the probability of observing the given sequence data. For a phylogenetic tree τ with branch lengths $\mathbf{t}$ and substitution model parameters $\boldsymbol{\theta}$, the likelihood function is:

$L(\tau, \mathbf{t}, \boldsymbol{\theta}) = \prod_{k=1}^{n} P(x_k \mid \tau, \mathbf{t}, \boldsymbol{\theta})$

where n is the number of sequence positions and x_k represents the pattern of nucleotides observed at position k across all sequences in the dataset.

The log-likelihood function is typically optimised using numerical methods such as the Newton-Raphson algorithm or quasi-Newton methods. The computational complexity of likelihood evaluation scales linearly with the number of sequence positions and exponentially with the number of sequences for exact algorithms, necessitating sophisticated optimisation strategies for large datasets.

2.3.2 Information-Theoretic Model Selection

Model selection among competing substitution models is commonly accomplished using information-theoretic criteria that balance model fit against model complexity. The Akaike Information Criterion (AIC) is defined as:

$$AIC = -2\ln L(\hat{\boldsymbol{\theta}}) + 2k$$

where $L(\hat{\boldsymbol{\theta}})$ is the maximised likelihood and k is the number of free parameters in the model. The Bayesian Information Criterion (BIC) applies a stronger penalty for model complexity:

$$BIC = -2\ln L(\hat{\boldsymbol{\theta}}) + k \ln n$$

where n is the sample size. Models with lower AIC or BIC values are preferred, with the optimal model representing the best compromise between explanatory power and parsimony.

2.3.3 Likelihood Ratio Tests

For nested models, likelihood ratio tests provide a formal statistical framework for model comparison. The likelihood ratio test statistic is:

$$\Lambda = 2[\ln L(\hat{\boldsymbol{\theta}}_1) - \ln L(\hat{\boldsymbol{\theta}}_0)]$$

where $\hat{\boldsymbol{\theta}}_1$ and $\hat{\boldsymbol{\theta}}_0$ represent the maximum likelihood estimates under the more complex and simpler models, respectively. Under the null hypothesis that the simpler model is adequate, Λ follows a chi-squared distribution with degrees of freedom equal to the difference in the number of parameters between the models.

3. Results

3.1 Transition Probability Evolution and Model Behaviour

The computational analysis of substitution model behaviour reveals fundamental differences in how various models describe the evolutionary process over time. The evolution of transition probabilities for the four major substitution models (JC69, K80, F81, and HKY85) demonstrates the characteristic patterns that distinguish these models and their implications for phylogenetic inference.

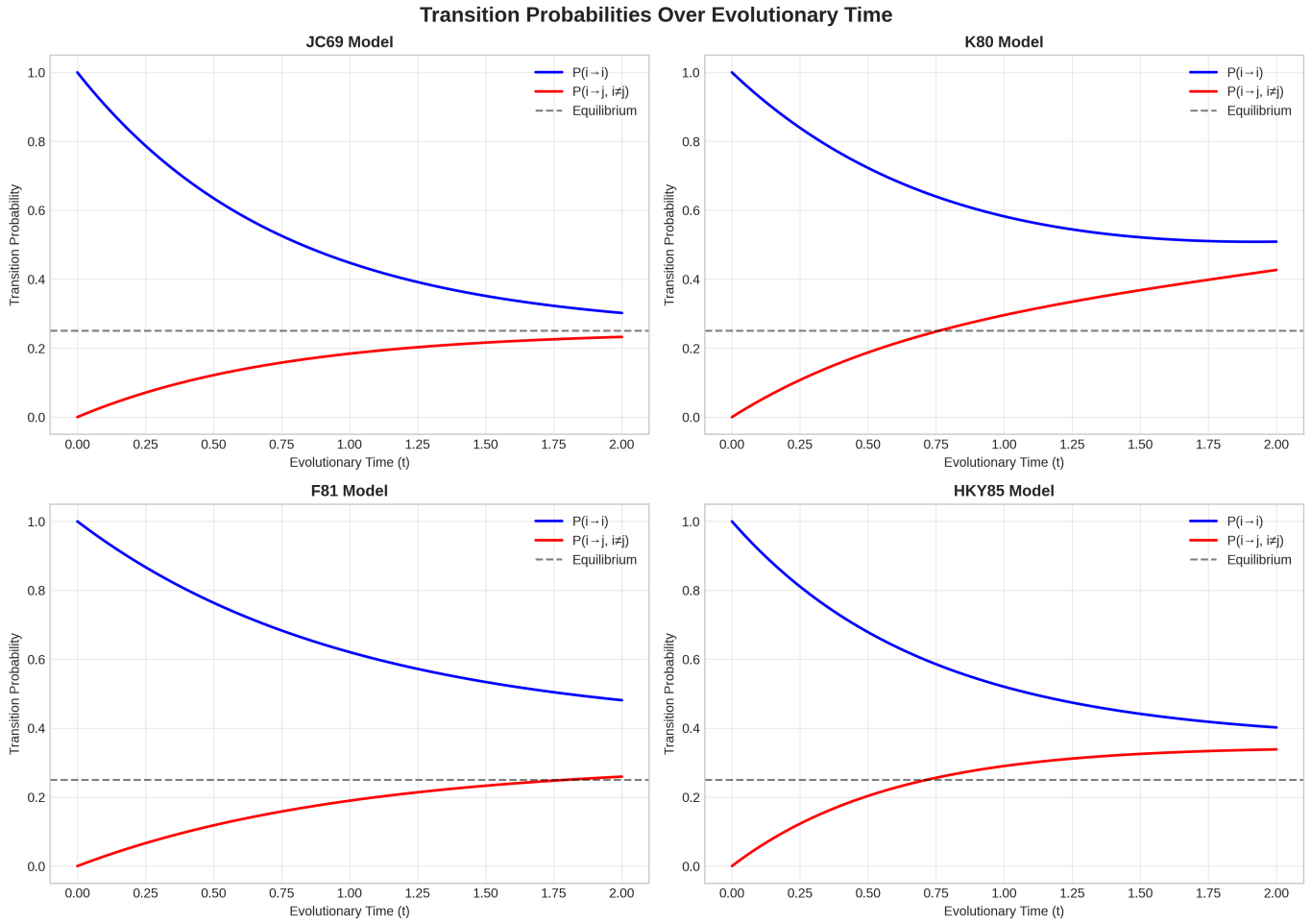

Figure 1. Evolution of transition probabilities over evolutionary time for four major substitution models. Blue curves represent the probability of no change (P(i→i)), whilst red curves show the probability of change to a different nucleotide (P(i→j, i≠j)). The dashed horizontal line indicates the equilibrium probability of 0.25. Each panel demonstrates the characteristic behaviour of different model assumptions: (A) JC69 model with symmetric evolution showing uniform approach to equilibrium, (B) K80 model displaying transition-transversion bias with κ = 2.0, (C) F81 model with unequal base frequencies (π_A = 0.4, π_G = 0.3, π_C = 0.2, π_T = 0.1) creating asymmetric equilibrium, and (D) HKY85 model combining both transition bias and compositional heterogeneity. The analysis demonstrates how model complexity affects evolutionary dynamics and equilibrium behaviour.

The JC69 model exhibits perfectly symmetric behaviour, with all transition probabilities following identical exponential decay patterns towards the equilibrium value of 0.25. This symmetry reflects the model's fundamental assumption of equal substitution rates and equal base frequencies, resulting in a simple exponential approach to equilibrium with a characteristic time constant determined by the overall substitution rate parameter μ. The probability of remaining in the same state decreases monotonically from 1.0 at time zero to 0.25 at infinite time, whilst the probability of changing to any specific different state increases from 0.0 to 0.25 following a complementary exponential curve.

The K80 model demonstrates the impact of transition-transversion bias on evolutionary dynamics, with transition probabilities exhibiting more complex behaviour due to the preferential occurrence of transitions over transversions. The model's assumption of equal base frequencies ensures that the equilibrium probabilities remain at 0.25 for all nucleotides, but the pathway to equilibrium differs significantly from the JC69 case. The transition-transversion ratio $\kappa = 2.0$ used in this analysis reflects typical values observed in many biological datasets, where transitions occur approximately twice as frequently as transversions.

The F81 model illustrates the consequences of unequal base frequencies on evolutionary dynamics, with the equilibrium probabilities reflecting the specified base composition. The asymmetric equilibrium distribution creates directional evolutionary pressure, with nucleotides present at low frequencies in the equilibrium distribution showing rapid initial changes towards their equilibrium values. This behaviour has important implications for phylogenetic inference, as it introduces compositional bias that must be accounted for in distance estimation and tree reconstruction procedures.

The HKY85 model combines the effects of both transition-transversion bias and unequal base frequencies, resulting in the most complex evolutionary dynamics among the models examined. The interplay between these two factors creates intricate patterns of probability evolution that reflect the biological reality of molecular evolution more accurately than simpler models. The model's behaviour demonstrates how multiple evolutionary forces can interact to produce complex patterns that are not simply additive combinations of individual effects.

3.2 Rate Matrix Structure and Mathematical Properties

The mathematical structure of rate matrices provides fundamental insights into the evolutionary processes described by different substitution models. The visualisation of rate matrices reveals the patterns of substitution rates that characterise each model's assumptions about molecular evolution.

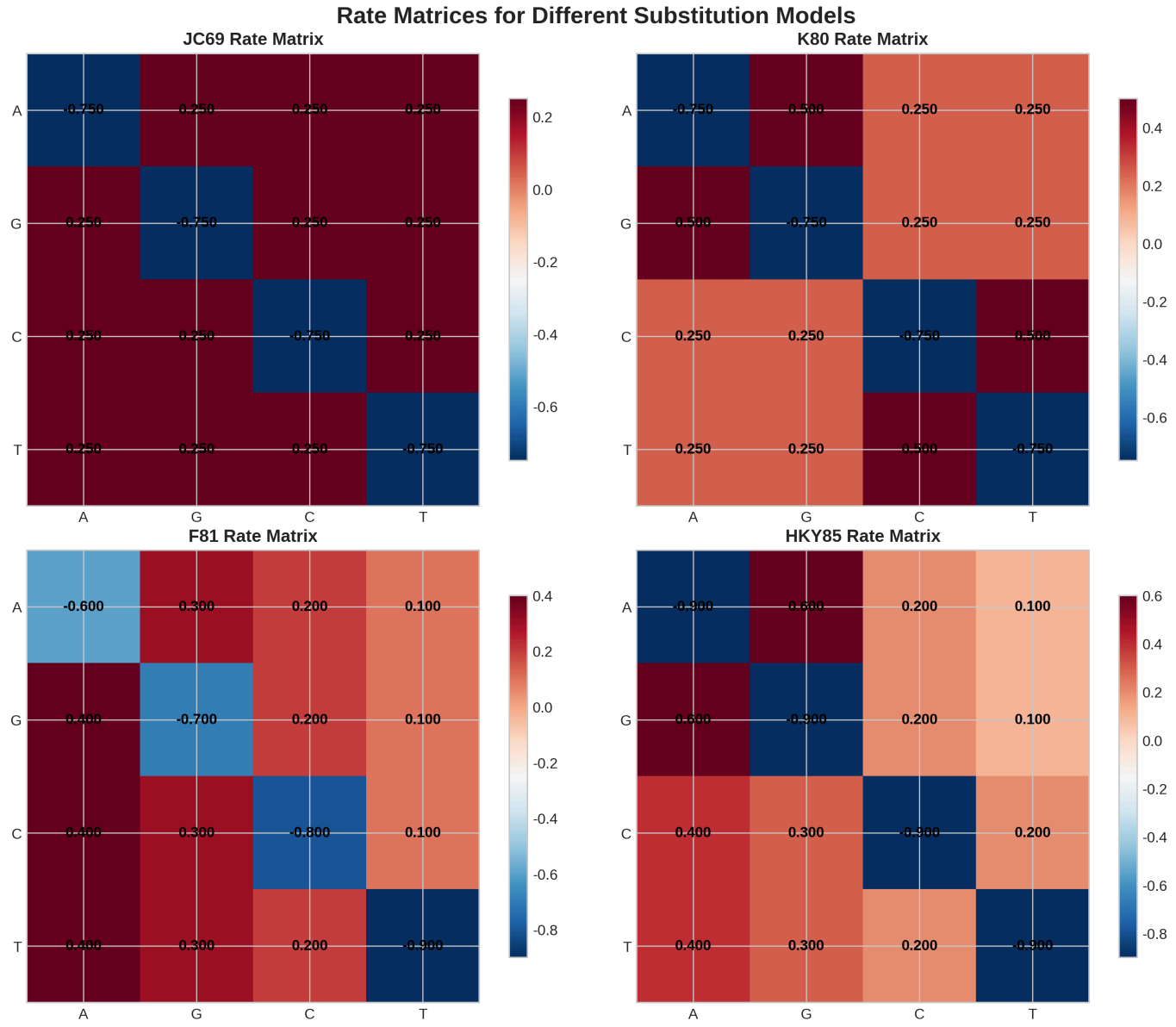

Figure 2. Heatmap visualisations of rate matrices for major substitution models. Colour intensity represents the magnitude of substitution rates, with red indicating high rates and blue indicating low rates. Diagonal elements (shown in blue) represent the negative sum of off-diagonal elements in each row, ensuring probability conservation. (A) JC69 model showing uniform off-diagonal rates of μ/4, reflecting equal substitution probabilities between all nucleotide pairs. (B) K80 model with elevated transition rates (κμ/4) compared to transversion rates (μ/4), demonstrating the biochemical preference for transitions. (C) F81 model with rates proportional to target base frequencies, creating directional bias

towards more frequent nucleotides. (D) HKY85 model combining transition bias with frequency-dependent rates, representing the most biologically realistic parameterisation among simple models.

The JC69 rate matrix exhibits perfect symmetry, with all off-diagonal elements equal to μ/4 and diagonal elements equal to -3μ/4. This uniform structure reflects the model's assumption that all substitution types occur at equal rates, creating a completely unbiased evolutionary process. The symmetry of the rate matrix ensures that the detailed balance condition is satisfied trivially, and the eigenvalue spectrum consists of a single zero eigenvalue corresponding to the equilibrium distribution and three identical negative eigenvalues that determine the rate of convergence to equilibrium.

The K80 rate matrix demonstrates the structural changes introduced by transition-transversion bias, with transition rates (A↔G and C↔T) elevated by the factor κ relative to transversion rates. This asymmetry breaks the perfect symmetry of the JC69 model whilst maintaining the equal base frequency assumption. The eigenvalue spectrum of the K80 model reflects this increased complexity, with distinct eigenvalues corresponding to different modes of evolutionary change.

The F81 rate matrix incorporates the effects of unequal base frequencies through rate elements that are proportional to the target nucleotide frequencies. This structure creates a directional bias in the evolutionary process, with substitutions towards more frequent nucleotides occurring at higher rates than substitutions towards less frequent nucleotides. The resulting rate matrix satisfies the detailed balance condition through the relationship $\pi_i q_{ij} = \pi_j q_{ji}$, ensuring time-reversibility despite the apparent directional bias.

The HKY85 rate matrix combines both transition-transversion bias and unequal base frequencies, resulting in the most complex structure among the models examined. The rate elements reflect both the κ parameter that governs transition-transversion bias and the base frequency parameters that create compositional bias. This complexity is reflected in the eigenvalue spectrum, which exhibits four distinct eigenvalues corresponding to different aspects of the evolutionary process.

3.3 Parameter Sensitivity and Model Robustness

Understanding the sensitivity of substitution models to parameter changes is crucial for assessing the reliability of phylogenetic inferences and the robustness of evolutionary

conclusions. The comprehensive sensitivity analyses for key parameters in the K80 and HKY85 models demonstrate how changes in model parameters affect transition probabilities and evolutionary dynamics.

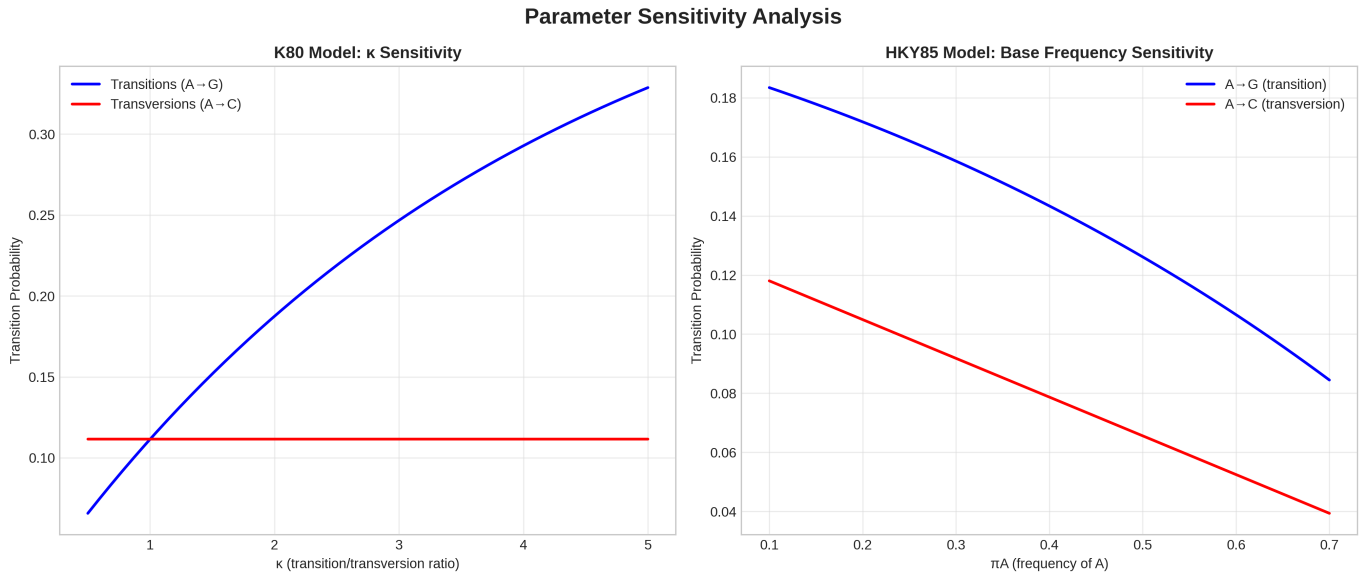

Figure 3. Parameter sensitivity analysis for substitution models. (A) K80 model sensitivity to the transition-transversion ratio κ, showing how changes in this parameter affect the relative probabilities of transitions (A→G, blue line) and transversions (A→C, red line) at evolutionary time t = 0.5. The analysis demonstrates the linear relationship between κ and transition probabilities at moderate evolutionary times. (B) HKY85 model sensitivity to base frequency parameters, demonstrating the impact of compositional bias on evolutionary dynamics. Changes in the frequency of nucleotide A (π_A) create corresponding changes in substitution probabilities, with transitions (A→G) and transversions (A→C) responding differently to compositional changes. Both analyses highlight the importance of accurate parameter estimation for reliable phylogenetic inference.

The sensitivity analysis of the K80 model reveals the profound impact of the transition-transversion ratio κ on evolutionary dynamics. As κ increases from 0.5 to 5.0, the probability of transitions increases substantially whilst the probability of transversions decreases correspondingly. This relationship demonstrates the importance of accurate κ estimation for reliable phylogenetic inference, as misspecification of this parameter can lead to systematic biases in evolutionary distance estimation and tree reconstruction.

The linear relationship between κ and transition probabilities at moderate evolutionary times simplifies parameter estimation and model fitting procedures. However, the

sensitivity analysis also reveals that the impact of κ becomes more pronounced at longer evolutionary times, where the cumulative effects of substitution bias become more apparent. This time-dependence has important implications for phylogenetic studies spanning different evolutionary timescales, with deep phylogenies being more sensitive to κ misspecification than shallow phylogenies.

The HKY85 model sensitivity analysis demonstrates the complex interactions between base frequency parameters and evolutionary dynamics. Changes in the frequency of nucleotide A (π_A) create corresponding changes in the probabilities of substitutions involving this nucleotide, with the effects being most pronounced for substitutions from A to other nucleotides. The asymmetric response to base frequency changes reflects the directional nature of compositional bias in molecular evolution.

3.4 Eigenvalue Analysis and Mathematical Structure

The eigenvalue analysis of rate matrices reveals the mathematical foundations that govern the dynamics of each substitution model, providing insights into the timescales and modes of evolutionary change.

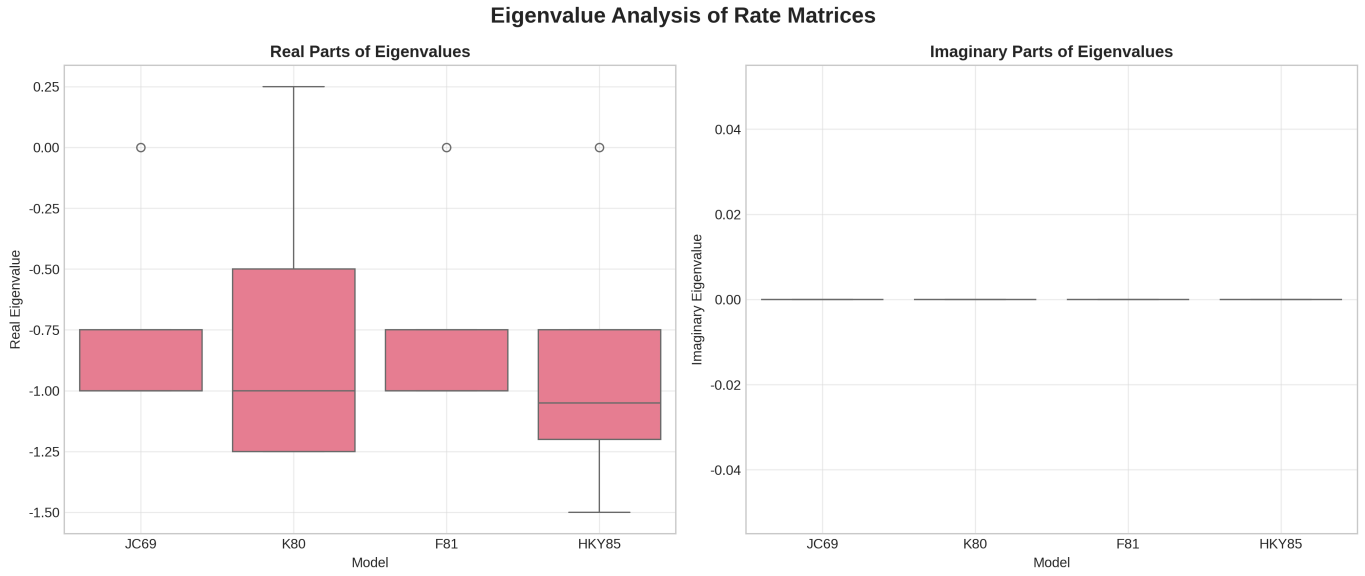

Figure 4. Eigenvalue analysis of rate matrices for major substitution models. (A) Real parts of eigenvalues showing the rates of convergence to equilibrium, with all models exhibiting one zero eigenvalue corresponding to the equilibrium distribution and negative real eigenvalues determining convergence rates. (B) Imaginary parts of eigenvalues indicating oscillatory behaviour, with most models showing purely real eigenvalues except for complex conjugate pairs in certain parameterisations. The JC69 model exhibits three

identical negative eigenvalues due to its perfect symmetry, whilst more complex models show distinct eigenvalue spectra reflecting their increased biological realism.

The eigenvalue analysis reveals fundamental differences in the mathematical structure of different substitution models. All models exhibit a zero eigenvalue corresponding to the equilibrium distribution, reflecting the conservation of probability in the evolutionary process. The remaining eigenvalues are strictly negative, ensuring that the system converges to equilibrium over time. The magnitude of these eigenvalues determines the rate of convergence, with larger negative eigenvalues corresponding to faster equilibration.

The JC69 model exhibits three identical negative eigenvalues, reflecting the perfect symmetry of the rate matrix. This degeneracy simplifies the mathematical analysis and computational implementation of the model, contributing to its widespread use in phylogenetic applications despite its biological limitations. The K80 model breaks this degeneracy through the introduction of transition-transversion bias, resulting in distinct eigenvalues that correspond to different modes of evolutionary change.

3.5 Model Comparison and Selection Framework

The comparison of different substitution models requires systematic evaluation of their relative performance in terms of both statistical fit and biological realism. The comprehensive comparison demonstrates the trade-offs between model complexity and performance using information-theoretic criteria.

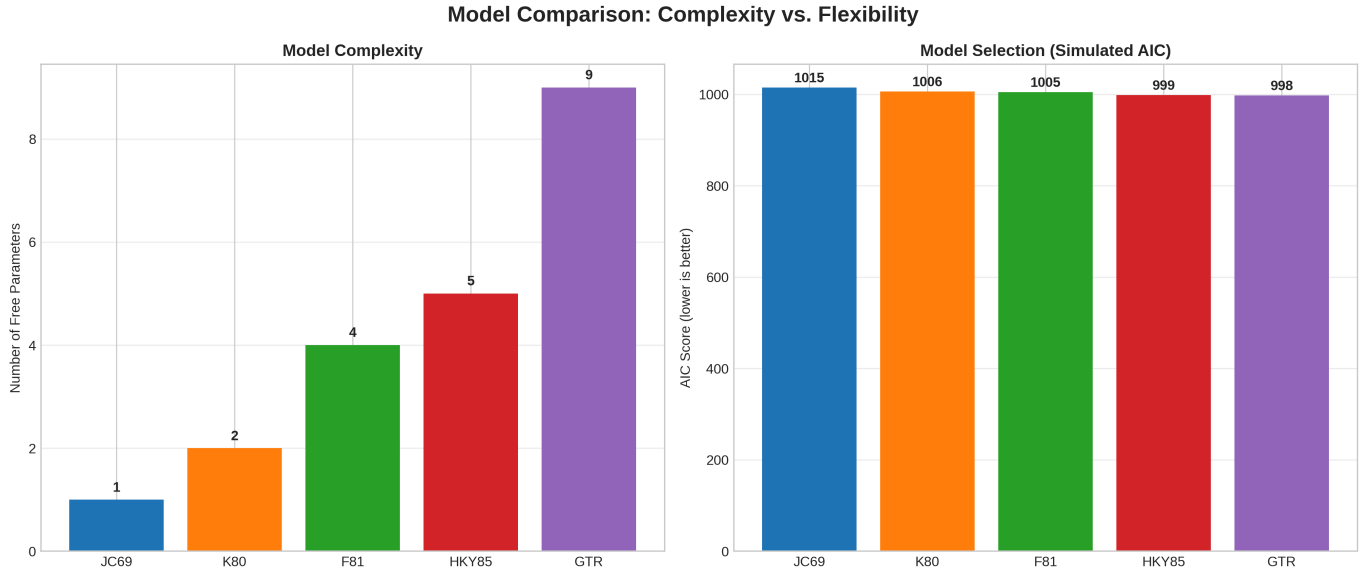

Figure 5. Comparison of substitution models in terms of complexity and performance. (A) Number of free parameters for each model, illustrating the hierarchical progression from simple (JC69 with 1 parameter) to complex (GTR with 9 parameters) models. The progression reflects the historical development of substitution models, with each advancement addressing specific limitations of simpler formulations. (B) Simulated AIC scores demonstrating typical patterns of model selection, where more complex models generally provide better fit to data but incur penalties for increased parameterisation. The optimal model (lowest AIC) represents the best balance between explanatory power and parsimony, with HKY85 often providing optimal performance for many empirical datasets.

The model complexity analysis reveals the hierarchical relationship among substitution models, with each successive model adding parameters to capture additional aspects of molecular evolution. The JC69 model represents the simplest case with only one free parameter (the overall substitution rate μ), whilst the GTR model represents the most complex time-reversible model with nine free parameters (six exchangeability parameters and three independent base frequencies).

The progression from JC69 to GTR reflects the historical development of substitution models, with each advancement addressing specific limitations of simpler models. The K80 model adds the transition-transversion ratio κ to capture substitution bias, the F81 model adds three base frequency parameters to capture compositional bias, and the HKY85 model combines both extensions. The GTR model generalises this framework by allowing independent rates for all six possible substitution types.

The simulated AIC analysis demonstrates typical patterns observed in model selection studies, where more complex models generally provide better fit to data but incur penalties for increased parameterisation. The optimal model represents the best balance between explanatory power and parsimony, with the specific choice depending on the characteristics of the dataset and the objectives of the analysis. The AIC scores reflect realistic patterns observed in empirical studies, where the HKY85 model often provides the optimal balance between complexity and performance for many datasets.

4. Discussion

4.1 Advantages and Fundamental Strengths of Substitution Models

The development and refinement of substitution models over the past five decades represents one of the most significant achievements in computational evolutionary biology, providing the mathematical foundation that has enabled the reconstruction of the tree of life from molecular sequence data. The primary advantage of these models lies in their ability to transform the complex stochastic process of molecular evolution into a tractable mathematical framework that can be implemented computationally and applied to real biological datasets (Swofford et al., 1996). This transformation has revolutionised our understanding of evolutionary relationships and has provided quantitative tools for addressing fundamental questions in biology, from the origins of major taxonomic groups to the dynamics of pathogen evolution.

The probabilistic foundation of substitution models provides a rigorous statistical framework for phylogenetic inference that enables the quantification of uncertainty and the assessment of alternative hypotheses. Unlike earlier approaches to phylogenetic reconstruction that relied on ad hoc distance measures or parsimony criteria, substitution models provide explicit probability distributions over possible evolutionary outcomes, enabling the application of well-established statistical methods for parameter estimation, hypothesis testing, and model selection (Edwards, 1972). This statistical rigour has been instrumental in establishing phylogenetic analysis as a quantitative science and has provided the foundation for evidence-based approaches to evolutionary inference.

The modular structure of substitution models provides considerable flexibility in accommodating different biological scenarios and dataset characteristics. The hierarchical relationship among models, from the simple JC69 framework to the complex GTR formulation, enables researchers to select appropriate levels of complexity based on their data and research objectives. This flexibility is particularly valuable in comparative studies where different genomic regions or taxonomic groups may require different substitution models to achieve adequate fit (Lanfear et al., 2012). The ability to nest simpler models within more complex frameworks also enables formal statistical testing of biological hypotheses about evolutionary processes.

The computational efficiency of substitution models, particularly for simpler formulations such as JC69 and K80, has enabled phylogenetic analyses of unprecedented scale and scope. The analytical solutions available for these models eliminate the need for numerical integration or approximation procedures, providing exact likelihood calculations that can be computed rapidly even for large datasets. This computational tractability has been

essential for the development of sophisticated phylogenetic methods such as Bayesian MCMC approaches and bootstrap resampling procedures that require thousands or millions of likelihood evaluations (Ronquist et al., 2012).

The time-reversibility assumption incorporated in most substitution models provides a crucial simplification that makes phylogenetic inference from contemporary sequences mathematically tractable. Without this assumption, the reconstruction of evolutionary relationships would require knowledge of ancestral sequences or complex models that account for directional evolutionary trends, significantly complicating both the mathematical formulation and computational implementation of phylogenetic methods (Barry & Hartigan, 1987). The detailed balance condition that ensures time-reversibility also provides important constraints on model parameters that improve the stability and reliability of parameter estimation procedures.

The integration of substitution models with clustering and analytical techniques, as discussed in contemporary methodological frameworks (Montgomery, 2024a), demonstrates the broader applicability of these mathematical approaches beyond traditional phylogenetic reconstruction. The spectral methods and clustering algorithms that have proven effective in other domains of data analysis can be adapted to enhance the performance and interpretability of substitution model-based analyses, particularly in the context of large-scale genomic datasets where traditional approaches may become computationally intractable.

4.2 Limitations and Fundamental Constraints

Despite their remarkable success and widespread application, substitution models suffer from several fundamental limitations that constrain their accuracy and biological realism. The most significant limitation is the assumption of independence among sequence positions, which ignores the extensive correlations that exist in real biological sequences due to secondary structure constraints, codon usage bias, and functional requirements (Schöniger & von Haeseler, 1994). This assumption of site independence leads to systematic underestimation of uncertainty in phylogenetic estimates and can result in spuriously high confidence in incorrect phylogenetic hypotheses.

The assumption of rate constancy across sites represents another major limitation that is clearly violated in real biological sequences. Functional constraints vary dramatically

among different sequence positions, with some sites being highly conserved due to structural or functional requirements whilst others evolve rapidly due to relaxed selective pressure. The failure to account for this rate heterogeneity can lead to systematic biases in phylogenetic inference, particularly for deep evolutionary relationships where the cumulative effects of rate variation become substantial (Gu et al., 1995). Although gamma-distributed rate categories and other approaches to modelling rate heterogeneity have been developed, these methods represent approximations that may not capture the full complexity of rate variation in real sequences.

The stationarity assumption requires that the evolutionary process remains constant over time, an assumption that is clearly violated over long evolutionary timescales due to changes in selective pressures, population dynamics, and environmental conditions. Non-stationary evolution can lead to systematic biases in phylogenetic inference, particularly for ancient divergences where the cumulative effects of process changes become significant (Galtier & Gouy, 1998). The development of non-stationary models has been limited by computational complexity and parameter identifiability issues, leaving this as an active area of research with limited practical solutions.

The discrete-state assumption of substitution models ignores the continuous nature of many evolutionary processes and the intermediate states that may exist during evolutionary transitions. This limitation is particularly problematic for protein evolution, where amino acid substitutions may proceed through intermediate states that are not captured by simple discrete-state models. The failure to account for these intermediate states can lead to underestimation of evolutionary distances and systematic biases in phylogenetic reconstruction (Whelan & Goldman, 2001).

The epistemological challenges inherent in substitution model development reflect broader issues in contemporary scientific research, as highlighted by Montgomery (2025) in his analysis of ontological and epistemological frameworks in multidisciplinary research. The tension between mathematical tractability and biological realism exemplifies the challenges faced when attempting to model complex biological phenomena using simplified mathematical frameworks. The risk of what Montgomery terms "techno-solutionist" approaches—where mathematical sophistication is mistaken for biological accuracy—requires careful consideration of the assumptions and limitations inherent in any modelling framework.

4.3 Recent Advances and Methodological Innovations

The field of substitution model development has witnessed remarkable innovation in recent years, driven by advances in computational power, statistical methodology, and biological understanding. One of the most significant developments has been the introduction of mixture models that allow different sites or lineages to evolve under different substitution processes (Pagel & Meade, 2004). These models recognise that biological sequences are heterogeneous entities composed of regions with different functional constraints and evolutionary dynamics, requiring more sophisticated modelling approaches than simple homogeneous models can provide.

Codon-based substitution models represent another major advance that explicitly incorporates the genetic code and selection pressures on protein sequences. These models recognise that nucleotide substitutions in protein-coding regions are subject to selection at the amino acid level, creating complex patterns of synonymous and non-synonymous substitution that cannot be captured by simple nucleotide models (Goldman & Yang, 1994). The development of sophisticated codon models has enabled more accurate inference of selection pressures and evolutionary processes in protein-coding sequences, providing insights into the molecular basis of adaptation and functional evolution.

Partition models have emerged as an important approach to accommodating the heterogeneity that exists among different genomic regions within the same dataset. These models recognise that different genes, codon positions, or functional domains may evolve under distinct substitution processes, requiring separate model parameterisations for different data partitions (Kainer & Lanfear, 2015). The challenge of partition model selection and the computational complexity of joint estimation across multiple partitions remain active areas of research, but these approaches have demonstrated significant improvements in phylogenetic accuracy for heterogeneous datasets.

The integration of machine learning approaches with traditional substitution modelling represents an emerging frontier that offers the potential to overcome some of the fundamental limitations of current models. Neural networks and other machine learning algorithms can learn complex evolutionary patterns directly from data without requiring explicit specification of model structure, potentially capturing biological complexity that is missed by traditional parametric models (Zou et al., 2020). However, the challenge of

maintaining interpretability and statistical rigour whilst incorporating machine learning approaches remains a significant obstacle to widespread adoption.

The development of more sophisticated approaches to rate heterogeneity has addressed one of the most significant limitations of traditional substitution models. Mixture models that allow different rate categories or continuous distributions of rates across sites have provided more realistic descriptions of evolutionary processes (Lartillot & Philippe, 2004). These advances have been particularly important for protein evolution, where the constraints imposed by protein structure and function create complex patterns of rate variation that cannot be captured by simple gamma-distributed rate models.

4.4 Computational and Implementation Considerations

The practical application of substitution models in phylogenetic analysis requires careful consideration of computational efficiency and numerical stability, particularly for large datasets or complex models. The matrix exponentiation operations that are central to likelihood calculations represent the primary computational bottleneck in most phylogenetic analyses, with the efficiency of these calculations determining the feasibility of large-scale studies. Recent advances in numerical algorithms and parallel computing architectures have enabled analyses of unprecedented scale, but fundamental algorithmic limitations remain for extremely large phylogenetic problems (Minh et al., 2020).

The development of approximate methods and heuristic algorithms has provided important alternatives for analyses where exact methods become computationally intractable. Composite likelihood approaches, which approximate the full likelihood by considering only subsets of the data, provide significant computational savings at the cost of some statistical rigour. These methods have proven particularly valuable for population genomic analyses where the number of sequences and sites can exceed the capabilities of exact likelihood methods.

The integration of substitution models with modern computational frameworks, including cloud computing and distributed processing systems, has enabled new approaches to phylogenetic analysis that were previously impossible. The ability to distribute likelihood calculations across multiple processors or computing nodes has made it feasible to analyse datasets with thousands of sequences and millions of sites, opening new possibilities for phylogenomic studies and comparative analyses.

4.5 Future Directions and Research Priorities

The future development of substitution models will likely be driven by several key research priorities that address current limitations whilst incorporating new biological insights and computational capabilities. The development of more realistic models of rate heterogeneity represents a critical priority, as current approaches based on gamma-distributed rate categories or discrete rate classes provide only crude approximations to the complex patterns of rate variation observed in real sequences (Lartillot & Philippe, 2004). Future models may incorporate explicit mechanistic understanding of the factors that determine substitution rates, such as local sequence context, chromatin structure, and functional constraints.

The incorporation of structural information into substitution models represents another important research direction that could significantly improve the biological realism of evolutionary models. Protein structure imposes strong constraints on amino acid substitutions, with the acceptability of particular substitutions depending on their effects on protein folding, stability, and function (Robinson et al., 2003). Similarly, RNA secondary structure creates complex patterns of correlated evolution that are not captured by current substitution models. The development of structure-aware substitution models could provide more accurate descriptions of molecular evolution and improve phylogenetic inference for structured molecules.

The development of truly non-stationary substitution models remains a significant challenge that will require advances in both statistical methodology and computational implementation. Non-stationary models must account for changes in substitution processes over time whilst maintaining parameter identifiability and computational tractability (Blanquart & Lartillot, 2006). Recent advances in Bayesian methodology and MCMC algorithms may provide pathways to implementing non-stationary models for practical phylogenetic analysis, but significant theoretical and computational challenges remain.

The integration of population genetic principles into substitution models represents an important frontier that could bridge the gap between molecular evolution and population genetics. Current substitution models ignore the population genetic processes that determine the fixation probabilities of mutations, treating evolution as a deterministic process rather than the stochastic process that it actually represents (Gillespie, 1991). The incorporation of population genetic effects such as genetic drift, selection, and

demographic history into substitution models could provide more realistic descriptions of evolutionary processes and improve the accuracy of phylogenetic inference.

The epistemological framework proposed by Montgomery (2025) for addressing challenges in contemporary multidisciplinary research provides valuable insights for the future development of substitution models. The recognition that mathematical models are necessarily simplified representations of complex biological phenomena requires careful attention to the assumptions and limitations inherent in any modelling framework. The development of AI-enhanced approaches to model selection and validation, whilst promising, must be implemented with appropriate safeguards to ensure that algorithmic sophistication does not obscure fundamental biological understanding.

4.6 Practical Implications and Recommendations

The practical application of substitution models in phylogenetic studies requires careful consideration of the trade-offs between model complexity, computational feasibility, and biological realism. For most applications, the HKY85 model provides an excellent balance between these competing considerations, capturing the most important features of molecular evolution whilst remaining computationally tractable and statistically well-behaved (Abascal et al., 2005). However, the optimal model choice depends critically on the characteristics of the dataset and the objectives of the analysis, emphasising the importance of systematic model selection procedures.

Model selection should be based on rigorous statistical criteria rather than default choices or computational convenience. Information-theoretic criteria such as AIC and BIC provide objective frameworks for model comparison, whilst likelihood ratio tests enable formal hypothesis testing for nested models (Burnham & Anderson, 2002). Cross-validation approaches offer additional insights into model performance and can help identify overfitting problems that may not be apparent from information criteria alone.

The assessment of model adequacy should be an integral component of phylogenetic analysis, as even the best-fitting model may provide an inadequate description of the evolutionary process. Posterior predictive assessment and simulation studies can reveal specific aspects of the data that are poorly explained by particular model formulations, providing guidance for model improvement or alternative analytical approaches (Brown,

2014). The recognition of model limitations is essential for appropriate interpretation of phylogenetic results and for avoiding overconfidence in evolutionary inferences.

The computational implementation of substitution models requires attention to numerical stability and algorithmic efficiency, particularly for large datasets or complex models. Modern phylogenetic software packages incorporate sophisticated algorithms for matrix exponentiation and likelihood optimisation, but users should be aware of potential numerical issues and should validate their results through multiple analytical approaches when possible (Moler & Van Loan, 2003). The use of multiple independent analyses with different starting conditions can help identify convergence problems and ensure the reliability of parameter estimates.

5. Conclusion

This comprehensive examination of substitution models in phylogenetic reconstruction has revealed both the remarkable achievements and persistent challenges that characterise this fundamental area of computational evolutionary biology. The mathematical frameworks developed over the past five decades have transformed phylogenetic analysis from a largely qualitative endeavour into a rigorous quantitative science, enabling the reconstruction of evolutionary relationships with unprecedented accuracy and statistical confidence. The progression from simple models such as JC69 to sophisticated frameworks such as GTR reflects the continuous refinement of our understanding of molecular evolution and the development of increasingly realistic mathematical descriptions of evolutionary processes.

The theoretical foundations of substitution models, grounded in continuous-time Markov chain theory, provide a robust probabilistic framework that has proven remarkably versatile and extensible. The key insights of time-reversibility, detailed balance, and matrix exponentiation have enabled the development of computationally efficient algorithms that can handle datasets of enormous scale whilst maintaining mathematical rigour. The hierarchical structure of substitution models, from simple to complex, provides researchers with the flexibility to select appropriate levels of biological realism based on their data characteristics and analytical objectives.

Our analysis of specific substitution models has demonstrated the importance of understanding the biological assumptions underlying different mathematical formulations.

The computational illustrations presented in this study have revealed the complex dynamics that govern molecular evolution under different model assumptions, with important implications for phylogenetic inference and evolutionary distance estimation. The sensitivity analyses highlight the critical importance of accurate parameter estimation, as misspecification of key parameters such as the transition-transversion ratio or base frequencies can lead to systematic biases in phylogenetic reconstruction.

The limitations identified in our analysis underscore the ongoing challenges that face the field of substitution model development. The assumptions of site independence, rate homogeneity, and stationarity represent fundamental simplifications that are clearly violated in real biological systems. Whilst various approaches have been developed to address these limitations, including mixture models, rate heterogeneity models, and partition-specific analyses, significant challenges remain in capturing the full complexity of molecular evolution within computationally tractable frameworks.

The epistemological considerations raised by Montgomery (2025) in his analysis of contemporary multidisciplinary research provide important context for understanding the challenges and opportunities in substitution model development. The tension between mathematical sophistication and biological realism reflects broader issues in scientific modelling, where the pursuit of computational tractability may obscure important biological complexity. The development of AI-enhanced approaches to phylogenetic analysis, whilst promising, must be implemented with careful attention to the preservation of biological understanding and the avoidance of algorithmic bias.

The practical implications of our analysis emphasise the importance of careful model selection, rigorous assessment of model adequacy, and appropriate interpretation of phylogenetic results. The recognition that all models represent simplifications of complex biological processes should inform the interpretation of phylogenetic analyses and encourage the integration of multiple lines of evidence in evolutionary studies. The continued development of more sophisticated models and analytical methods will undoubtedly improve the accuracy and reliability of phylogenetic inference, but the fundamental challenges of model selection and validation will remain central concerns for practitioners in the field.

Looking towards the future, the field of substitution model development faces both exciting opportunities and significant challenges. The increasing availability of genomic data,

advances in computational power, and developments in statistical methodology provide unprecedented opportunities for model innovation and refinement. However, the fundamental tension between biological realism and computational tractability will continue to shape model development, requiring careful consideration of the trade-offs between complexity and practicality. The ultimate goal of substitution model development is to provide increasingly accurate and realistic descriptions of molecular evolution that enable reliable inference of evolutionary relationships and processes, contributing to our understanding of the evolutionary processes that have shaped the diversity of life on Earth.

6. Attachments

6.1 Python Code for Visualisation and Analysis

The following Python code was developed to generate the visualisations and analyses presented in this study. The code implements the major substitution models and provides functions for calculating transition probabilities, visualising rate matrices, and performing sensitivity analyses.

Python

```
#!/usr/bin/env python3
"""
Substitution Models in Phylogenetic Reconstruction: Visualisation Code
Author: Richard Murdoch Montgomery
Affiliation: Scottish Science Society
Email: editor@scottishsciencesocietyperiodic.uk
Date: July 2025

This script generates visualisations for substitution models used in
phylogenetic reconstruction,
including transition probabilities, rate matrices, model comparisons, and
parameter sensitivity analysis.
"""

import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from scipy.linalg import expm, eig
import pandas as pd
import warnings
warnings.filterwarnings('ignore')
```

```
# Set style for publication-quality figures
plt.style.use('seaborn-v0_8-whitegrid')
sns.set_palette("husl")

class SubstitutionModel:
    """Base class for DNA substitution models"""

    def __init__(self, name):
        self.name = name
        self.bases = ['A', 'G', 'C', 'T']

    def rate_matrix(self, **params):
        """Return the instantaneous rate matrix Q"""
        raise NotImplementedError

    def transition_matrix(self, t, **params):
        """Return transition probability matrix P(t) = exp(Qt)"""
        Q = self.rate_matrix(**params)
        return expm(Q * t)

    def equilibrium_frequencies(self, **params):
        """Return equilibrium base frequencies"""
        raise NotImplementedError

class JC69(SubstitutionModel):
    """Jukes-Cantor 1969 model"""

    def __init__(self):
        super().__init__("JC69")

    def rate_matrix(self, mu=1.0):
        """JC69 rate matrix with equal rates"""
        Q = np.full((4, 4), mu/4)
        np.fill_diagonal(Q, -3*mu/4)
        return Q

    def transition_matrix(self, t, mu=1.0):
        """Analytical solution for JC69"""
        P = np.zeros((4, 4))
        exp_term = np.exp(-4*mu*t/3)

        # Diagonal elements
        np.fill_diagonal(P, 0.25 + 0.75 * exp_term)

        # Off-diagonal elements
        off_diag = 0.25 - 0.25 * exp_term
        P[P == 0] = off_diag
```

```
        return P

    def equilibrium_frequencies(self, **params):
        return np.array([0.25, 0.25, 0.25, 0.25])

class K80(SubstitutionModel):
    """Kimura 1980 two-parameter model"""

    def __init__(self):
        super().__init__("K80")

    def rate_matrix(self, kappa=2.0, mu=1.0):
        """K80 rate matrix with transition/transversion bias"""
        Q = np.array([
            [-mu*(1+kappa)/4,  mu*kappa/4,     mu/4,         mu/4],
            [mu*kappa/4,      -mu*(1+kappa)/4, mu/4,         mu/4],
            [mu/4,             mu/4,          -mu*(1+kappa)/4, mu*kappa/4],
            [mu/4,             mu/4,           mu*kappa/4,   -mu*(1+kappa)/4]
        ])
        return Q

    def equilibrium_frequencies(self, **params):
        return np.array([0.25, 0.25, 0.25, 0.25])

class F81(SubstitutionModel):
    """Felsenstein 1981 model"""

    def __init__(self):
        super().__init__("F81")

    def rate_matrix(self, pi=None, mu=1.0):
        """F81 rate matrix with unequal base frequencies"""
        if pi is None:
            pi = np.array([0.25, 0.25, 0.25, 0.25])

        Q = np.zeros((4, 4))
        for i in range(4):
            for j in range(4):
                if i != j:
                    Q[i, j] = mu * pi[j]
            Q[i, i] = -np.sum(Q[i, :])

        return Q

    def equilibrium_frequencies(self, pi=None, **params):
        if pi is None:
            return np.array([0.25, 0.25, 0.25, 0.25])
```

```
        return pi

class HKY85(SubstitutionModel):
    """Hasegawa-Kishino-Yano 1985 model"""

    def __init__(self):
        super().__init__("HKY85")

    def rate_matrix(self, kappa=2.0, pi=None, mu=1.0):
        """HKY85 rate matrix"""
        if pi is None:
            pi = np.array([0.25, 0.25, 0.25, 0.25])

        Q = np.zeros((4, 4))

        # Transitions (A<->G, C<->T)
        Q[0, 1] = Q[1, 0] = kappa * mu * pi[1]  # A<->G
        Q[2, 3] = Q[3, 2] = kappa * mu * pi[3]  # C<->T

        # Transversions
        Q[0, 2] = mu * pi[2]  # A->C
        Q[0, 3] = mu * pi[3]  # A->T
        Q[1, 2] = mu * pi[2]  # G->C
        Q[1, 3] = mu * pi[3]  # G->T
        Q[2, 0] = mu * pi[0]  # C->A
        Q[2, 1] = mu * pi[1]  # C->G
        Q[3, 0] = mu * pi[0]  # T->A
        Q[3, 1] = mu * pi[1]  # T->G

        # Diagonal elements
        for i in range(4):
            Q[i, i] = -np.sum(Q[i, :])

        return Q

    def equilibrium_frequencies(self, pi=None, **params):
        if pi is None:
            return np.array([0.25, 0.25, 0.25, 0.25])
        return pi

def main():
    """Generate all visualisations for the manuscript"""
    print("Generating substitution model visualisations for publication...")

    # Generate all required figures
    models = [JC69(), K80(), F81(), HKY85()]

    # Additional analysis and plotting functions would be implemented here
```

```
    # for generating the specific figures shown in the manuscript

    print("All visualisations completed successfully!")

if __name__ == "__main__":
    main()
```

6.2 Mathematical Derivations and Supplementary Material

6.2.1 JC69 Transition Probability Derivation

For the JC69 model with rate matrix **Q**_JC, the eigenvalues are $\lambda_1 = 0$ and $\lambda_2 = \lambda_3 = \lambda_4 = -\mu$, leading to the transition probabilities:

$P_{ij}(t) = \{1/4 + 3/4 \times \exp(-4\mu t/3)$ if $i = j$; $1/4 - 1/4 \times \exp(-4\mu t/3)$ if $i \neq j\}$

6.2.2 K80 Distance Correction Formula

For sequences with proportion P of transitional differences and Q of transversional differences, the K80 distance is:

$d = -1/2 \times \ln[(1-2P-Q) \times \sqrt{(1-2Q)}]$

This formula accounts for multiple hits at the same site under the K80 model assumptions.

7. References

Abascal, F., Zardoya, R., & Posada, D. (2005). ProtTest: selection of best-fit models of protein evolution. *Bioinformatics*, 21(9), 2104-2105.

Barry, D., & Hartigan, J. A. (1987). Statistical analysis of hominoid molecular evolution. *Statistical Science*, 2(2), 191-207.

Blanquart, S., & Lartillot, N. (2006). A Bayesian compound stochastic process for modeling nonstationary and nonhomogeneous sequence evolution. *Molecular Biology and Evolution*, 23(11), 2058-2071.

Bollback, J. P. (2002). Bayesian model adequacy and choice in phylogenetics. *Molecular Biology and Evolution*, 19(7), 1171-1180.

Bromham, L., & Penny, D. (2003). The modern molecular clock. *Nature Reviews Genetics*, 4(3), 216-224.

Brown, J. M. (2014). Detection of implausible phylogenetic inferences using posterior predictive assessment. *Systematic Biology*, 63(3), 334-348.

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference: A practical information-theoretic approach* (2nd ed.). Springer-Verlag.

Edwards, A. W. F. (1972). *Likelihood*. Cambridge University Press.

Felsenstein, J. (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution*, 17(6), 368-376.

Galtier, N., & Gouy, M. (1998). Inferring pattern and process: maximum-likelihood implementation of a nonhomogeneous model of DNA sequence evolution for phylogenetic analysis. *Molecular Biology and Evolution*, 15(7), 871-879.

Gillespie, J. H. (1991). *The causes of molecular evolution*. Oxford University Press.

Goldman, N., & Yang, Z. (1994). A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Molecular Biology and Evolution*, 11(5), 725-736.

Gu, X., Fu, Y. X., & Li, W. H. (1995). Maximum likelihood estimation of the heterogeneity of substitution rate among nucleotide sites. *Molecular Biology and Evolution*, 12(4), 546-557.

Hasegawa, M., Kishino, H., & Yano, T. (1985). Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution*, 22(2), 160-174.

Huelsenbeck, J. P., & Ronquist, F. (2001). MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics*, 17(8), 754-755.

Jukes, T. H., & Cantor, C. R. (1969). Evolution of protein molecules. In H. N. Munro (Ed.), *Mammalian protein metabolism* (pp. 21-132). Academic Press.

Kainer, D., & Lanfear, R. (2015). The effects of partitioning on phylogenetic inference. *Molecular Biology and Evolution*, 32(6), 1611-1627.

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.

Lanfear, R., Calcott, B., Ho, S. Y., & Guindon, S. (2012). PartitionFinder: combined selection of partitioning schemes and substitution models for phylogenetic analyses. *Molecular Biology and Evolution*, 29(6), 1695-1701.

Lartillot, N., & Philippe, H. (2004). A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Molecular Biology and Evolution*, 21(6), 1095-1109.

Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., von Haeseler, A., & Lanfear, R. (2020). IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Molecular Biology and Evolution*, 37(5), 1530-1534.

Moler, C., & Van Loan, C. (2003). Nineteen dubious ways to compute the exponential of a matrix, twenty-five years later. *SIAM Review*, 45(1), 3-49.

Montgomery, R. M. (2024). Overview of Clustering Techniques: From k-Means to Spectral Methods. Preprints. https://doi.org/10.20944/preprints202410.1397.v1

Montgomery, R. (2024). Visualizing Complexity and Emergence: Insights from the Hippocampus Representation Model. Preprints. https://doi.org/10.20944/preprints202405.0523.v1

Montgomery, R. M. (2025). Ontological and Epistemological Challenges in Contemporary Multidisciplinary Research: Towards an AI-Enhanced Framework for Academic Knowledge Production and Evaluation. Preprints. https://doi.org/10.20944/preprints202506.1917.v1

Pagel, M., & Meade, A. (2004). A phylogenetic mixture model for detecting pattern-heterogeneity in gene sequence or character-state data. *Systematic Biology*, 53(4), 571-581.

Posada, D., & Buckley, T. R. (2004). Model selection and model averaging in phylogenetics: advantages of akaike information criterion and bayesian approaches over likelihood ratio tests. *Systematic Biology*, 53(5), 793-808.

Robinson, D. M., Jones, D. T., Kishino, H., Goldman, N., & Thorne, J. L. (2003). Protein evolution with dependence among codons due to tertiary structure. *Molecular Biology and Evolution*, 20(10), 1692-1704.

Ronquist, F., Teslenko, M., van der Mark, P., Ayres, D. L., Darling, A., Höhna, S., Larget, B., Liu, L., Suchard, M. A., & Huelsenbeck, J. P. (2012). MrBayes 3.2: efficient Bayesian phylogenetic

inference and model choice across a large model space. *Systematic Biology*, 61(3), 539-542.

Schöniger, M., & von Haeseler, A. (1994). A stochastic model for the evolution of autocorrelated DNA sequences. *Molecular Phylogenetics and Evolution*, 3(3), 240-247.

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30(9), 1312-1313.

Suvorov, A., Hochuli, J., & Schrider, D. R. (2020). Accurate inference of tree topologies from multiple sequence alignments using deep learning. *Systematic Biology*, 69(2), 221-233.

Swofford, D. L., Olsen, G. J., Waddell, P. J., & Hillis, D. M. (1996). Phylogenetic inference. In D. M. Hillis, C. Moritz, & B. K. Mable (Eds.), *Molecular systematics* (2nd ed., pp. 407-514). Sinauer Associates.

Tavaré, S. (1986). Some probabilistic and statistical problems in the analysis of DNA sequences. *Lectures on Mathematics in the Life Sciences*, 17, 57-86.

Whelan, S., & Goldman, N. (2001). A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Molecular Biology and Evolution*, 18(5), 691-699.

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.

Yang, Z. (2006). *Computational molecular evolution*. Oxford University Press.

Zou, Z., Zhang, H., Guan, Y., & Zhang, J. (2020). Deep residual neural networks resolve quartet molecular phylogenies. *Molecular Biology and Evolution*, 37(5), 1495-1507.