A novel molecular clock model based on anomalous diffusion

Lucas Goiriz[1,2], Raúl Ruiz[1] Òscar Garibo-i-Orts[2], J. Alberto Conejero[2], and Guillermo Rodrigo[1,†]

[1] Institute for Integrative Systems Biology (I2SysBio), CSIC – Universitat de València, 46980 Paterna, Spain
[†] guillermo.rodrigo@csic.es,
WWW home page: https://biosysdesign.csic.es
[2] Institute for Pure and Applied Mathematics (IUMPA), Universitat Politècnica de València, 46022 Valencia, Spain

Abstract. **Keywords:** anomalous diffusion, dynamic systems, virus evolution rate, stochastic process

1 The Molecular Clock Hypothesis

The molecular clock hypothesis, introduced by Emile Zuckerkandl and Linus Pauling in 1965, had a profound impact on evolutionary biology, becoming a vital tool for understanding and dating evolutionary events across various life forms, including archaea, eukarya, bacteria, and even viruses. This hypothesis proposes that the rate of genetic mutations in DNA sequences remains relatively steady over time, resembling a continuous "ticking clock", enabling researchers to estimate species divergence and evolutionary histories [1]. Motoo Kimura further developed this theory into a comprehensive framework in 1968, known as the neutral theory of molecular evolution. According to Kimura's proposal, the rate of molecular evolution is determined by the fixation rate of neutral mutations, which are not influenced by natural selection [2]. This theory also predicts that molecular clocks behave like Poissonian point processes [3, 4].

The application of the molecular clock hypothesis and its framework has been extensive in a wide range of biological problems, including estimating divergence times between species, reconstructing evolutionary relationships, calibrating phylogenetic trees, and tracking disease transmission and epidemics [5–7].

Despite its widespread use, the molecular clock hypothesis has been a subject of intense debate within the scientific community as the process of evolution is complex and affected by environmental changes, transmission bottlenecks, recombination, and speciation events, making it a highly volatile and stochastic phenomenon. Some studies have even shown that the molecular clock model is not valid in numerous cases [8], and others argue that it may not be applicable to all species and populations [9, 10].

Given these discrepancies, several modified molecular clock models have been proposed to address specific complexities and challenges in evolutionary studies.

These models include relaxed molecular clocks, Bayesian molecular clocks, birth-death molecular clocks, and relaxed clocks with covariates [11, 12], among others. These adaptations build upon the foundational ideas of the original molecular clock hypothesis while accommodating real-world complexities and expanding the scope of evolutionary research.

To motivate the development of a novel molecular clock model, we will briefly explore how a Poisson point process can be employed to model DNA sequence evolution. We will then extend the Poisson point process into a continuous stochastic process.

1.1 Evolution as a Poisson point process

The Poisson distribution is commonly used to model the occurrence of infrequent events within a fixed time or space interval. In the context of genetic mutations during DNA replication, each generation (defined as a replicative cycle) can introduce changes or substitutions in the DNA sequence due to various factors like errors induced by the DNA polymerase, radiation, or chemicals. Since the likelihood of a mutation at a specific position in the DNA sequence is assumed to be small, constant, and independent between generations (as supported by experimental evidence), the number of mutations in a lineage over n generations can be accurately described using the Poisson distribution.

Let u be the rate of mutations per generation, and n the number of generations. In this scenario, the number of mutations that occur in a lineage during these n generations follows a Poisson distribution with a mean value of un. In addition, if each generation takes the same amount of time, the number of mutations in the lineage during a specific time period t can be described as a homogeneous Poisson point process, denoted as $\{N(t), t \geq 0\}$, where $N(t)$ represents the total number of mutations that have taken place up to (and including) time t. Consequently, the probability of observing exactly n mutations, denoted as $N(t) = n$ at time t, is given by

$$\Pr(N(t) = n) = \frac{e^{-\kappa t} (\kappa t)^n}{n!} \tag{1}$$

where κ the rate of substitutions for a given unit of time. Importantly, **Equation 1** implies that the number of mutations in a lineage at time $t = 0$ is 0 and that the increments of the process are independent.

For further developments, it is convenient to compute the moment generating function, $M_{N(t)}(s)$, of the Poisson process:

$$
\begin{aligned}
M_{N(t)}(s) &= \mathbb{E}\left[e^{tN(t)}\right] && (2)\\
&= \sum_{v=0}^{\infty} e^{sn} \frac{e^{-\kappa t}\left(\kappa t\right)^n}{n!}\\
&= e^{-\kappa t} \sum_{n=0}^{\infty} \frac{\left(\kappa t e^s\right)^n}{n!}\\
&= e^{-\kappa t} e^{\kappa t e^s}\\
&= e^{\kappa t (e^s - 1)} && (3)
\end{aligned}
$$

This way it is trivial to demonstrate that the mean and variance of the process are both given by κt:

$$
\begin{aligned}
\mathbb{E}\left[N(t)\right] &= \frac{\partial}{\partial s} M_{N(t)}(s)\Big|_{s=0} && (4)\\
&= \left[\kappa t e^{\kappa t (e^s - 1) + s}\right]_{s=0}\\
&= \kappa t && (5)\\
\mathbb{V}\left[N(t)\right] &= \mathbb{E}\left[\left(N(t) - \mathbb{E}\left[N(t)\right]\right)^2\right] && (6)\\
&= \mathbb{E}\left[N^2(t) - 2N(t)\mathbb{E}\left[N(t)\right] + \mathbb{E}\left[N(t)\right]^2\right]\\
&= \mathbb{E}\left[N^2(t)\right] - \mathbb{E}\left[N(t)\right]^2\\
&= \frac{\partial^2}{\partial s^2} M_{N(t)}(s)\Big|_{s=0} - \left(\kappa t\right)^2\\
&= \left[\kappa t \left(\kappa t e^s + 1\right) e^{\kappa t (e^s - 1) + s}\right]_{s=0} - \left(\kappa t\right)^2\\
&= \left(\kappa t\right)^2 + \kappa t - \left(\kappa t\right)^2\\
&= \kappa t && (7)
\end{aligned}
$$

As a corollary, it is trivial to assess that the process' dispersion index $\rho_{N(t)}$, defined as the ratio between mean and variance, is equal to 1.

1.2 Evolution approximated as a continuous stochastic process

Similar to how the Poisson distribution can be approximated by a Gaussian distribution through the central limit theorem, a Poisson point process can also be approximated by a Wiener process. The Wiener process, also known as Brownian motion, is a continuous-time stochastic process characterized by independent and stationary increments. It is usually represented as $\{W(t), t \geq 0\}$, where $W(t)$ is a random variable representing the displacement of a particle at time t, its increments follow a normal distribution with a mean $\mathbb{E}[W(t)] = 0$ and, if it's the

standard Wiener process, a variance $\mathbb{V}[W(t)] = 1$. The Wiener process is widely used as a model for random fluctuations in various physical systems.

Therefore, the number of mutations during DNA replication can be reformulated as the following Langevin stochastic differential equation

$$\frac{dm(t)}{dt} = \kappa + \sqrt{\kappa}\xi(t) \tag{8}$$

where $\xi(t)$ is a Gaussian white noise characterized by $\mathbb{E}[\xi(t)] = 0$ and covariance function $\mathrm{Cov}[\xi(t)\xi(s)] = \delta(t-s)$. Note that $\xi(t)$ is defined as the formal derivative of the standard Wiener process $W(t)$, an assertion which has to be handled with caution because the Wiener process is nowhere differentiable with probability 1. **Equation 8** can be solved analytically:

$$\begin{aligned}
\frac{dm(t)}{dt} &= \kappa + \sqrt{\kappa}\xi(t) \\
m(t) &= m(0) + \kappa t + \sqrt{\kappa}\int_0^t \xi(s)ds \\
&= \kappa t + \sqrt{\kappa}\int_0^t \xi(s)ds
\end{aligned} \tag{9}$$

Note that this reformulation maintains the Poisson process' mean and variance:

$$\begin{aligned}
\mathbb{E}[m(t)] &= \mathbb{E}\left[\kappa t + \sqrt{\kappa}\int_0^t \xi(s)ds\right] && (10)\\
&= \kappa t + \sqrt{\kappa}\int_0^t \mathbb{E}[\xi(s)]\,ds \\
&= \kappa t && (11)\\
\mathbb{V}[m(t)] &= \mathbb{E}\left[(m(t) - \mathbb{E}[m(t)])^2\right] && (12)\\
&= \mathbb{E}\left[\left(\kappa t + \sqrt{\kappa}\int_0^t \xi(s)ds - \kappa t\right)^2\right] \\
&= \kappa\int_0^t\int_0^t \mathbb{E}[\xi(s)\xi(u)]\,dsdu \\
&= \kappa\int_0^t\int_0^t \delta(s-u)dsdu \\
&= \kappa\int_0^t 1du \\
&= \kappa t && (13)
\end{aligned}$$

As a corollary, the corresponding dispersion index $\rho_{m(t)}$ remains equal to 1 as expected, since the Wiener process is a continuous-time approximation of

the Poisson process. One concern that arises from this reformulation is that the number of mutations $m(t)$ is no longer an integer. However, this issue can be easily solved by applying a rounding function to $m(t)$ whenever it is necessary to obtain an integer value.

2 Anomalous Diffusion

In the preceding section we demonstrated that, according to the molecular clock hypothesis, the number of mutations occurring in a lineage during a specific time period t can be described as a Brownian motion exhibiting a mean and variance equal to κt, where κ represents the rate of substitutions for a given unit of time, akin to a microscopic particle moving in a fluid as a consequence of thermal forces.

However, it is well known that the diffusion of microscopic particles in a fluid does not always conform to Brownian motion. In fact, the diffusion of particles in a fluid can be classified into three main categories depending on their mean squared displacement (MSD; also understood as the variance of the stochastic process governing the motion): normal diffusion, subdiffusion, and superdiffusion. Under normal diffusion, the MSD of the particle is proportional to t, while under subdiffusion and superdiffusion the MSD of the particle is proportional to t^α, where α is known as the diffusion exponent, with $\alpha < 1$ for the former case and $\alpha > 1$ for the latter [13].

Similarly to a microscopic particle moving in a fluid, the number of mutations in a lineage during a specific time period t may not conform to a Brownian motion, as described by several studies observing overdispersed and underdispersed populations. Therefore, it is reasonable to consider that the number of mutations in a lineage during a specific time period t may exhibit anomalous diffusion.

2.1 Evolution as a fractional Brownian motion

Multiple stochastic definitions of anomalous diffusion exist, and it is usually left to the researcher to use the one that best fits their problem. In this work, fractional Brownian motion (fBm) is used as a model for anomalous diffusion due to its simple yet powerful mathematical properties.

fBm is a continuous-time stochastic process, represented as $\{W_\alpha(t), t \geq 0\}$, where $W_\alpha(t)$ is a random variable representing the displacement of a particle at time t, characterized by stationary increments, mean $\mathbb{E}[W_\alpha(t)] = 0$ and a covariance function of the form $\mathrm{Cov}[W_\alpha(t)W_\alpha(s)] = \frac{1}{2}(t^\alpha + s^\alpha - |t-s|^\alpha)$, where α is the diffusion exponent, which determines the degree of long-term dependence of the process. Indeed, the fBm is a generalization of the Wiener process, which corresponds to the case $\alpha = 1$.

To reformulate the number of mutations in a lineage during a specific time period t as a fBm, we will modify the Langevin stochastic differential equation shown in **Equation 8**:

$$\frac{dm(t)}{dt} = \kappa + \sqrt{\kappa}\eta(t) \tag{14}$$

where $\eta(t)$ is an appropriate noise source characterized by $\mathbb{E}[\eta(t)] = 0$ and a covariance function such that $\mathrm{Cov}[W_\alpha(t)W_\alpha(s)] = \int_0^t \int_0^s \mathrm{Cov}[\eta(u)\eta(v)]\,dudv$. It is trivial to compute that $\mathrm{Cov}[\eta(t)\eta(s)] = \frac{\alpha}{2}(\alpha-1)\,|t-s|^{\alpha-2}$. This definition allows for the computation of the appropriate mean and variance of the process:

$$\begin{aligned}
\mathbb{E}[m(t)] &= \mathbb{E}\left[\kappa t + \sqrt{\kappa}\int_0^t \eta(s)ds\right] && (15)\\
&= \kappa t && (16)\\
\mathbb{V}[m(t)] &= \mathbb{E}\left[(m(t) - \mathbb{E}[m(t)])^2\right] && (17)\\
&= \kappa\mathbb{E}\left[\left(\int_0^t \eta(s)ds\right)^2\right]\\
&= \kappa\int_0^t\int_0^t \mathbb{E}[\eta(s)\eta(u)]\,dsdu\\
&= \frac{\alpha\kappa}{2}(\alpha-1)\int_0^t\int_0^t |s-u|^{\alpha-2}\,dsdu\\
&= \frac{\alpha\kappa}{2}(\alpha-1)\int_0^t\left[\int_0^u (u-s)^{\alpha-2}\,ds + \int_u^t (s-u)^{\alpha-2}\,ds\right]du\\
&= \frac{\alpha\kappa}{2}\int_0^t \left[s^{\alpha-1} + (t-s)^{\alpha-1}\right]du\\
&= \kappa t^\alpha && (18)
\end{aligned}$$

Therefore, by using fBm as a model for anomalous diffusion, the number of mutations in a lineage during a specific time period t can be described as a stochastic process with a mean and variance equal to κt (in line with the molecular clock hypothesis) and κt^α, respectively. These values provide valuable insights into the reasons behind overdispersed and underdispersed genetic populations and the extent of their long-term dependence. It is essential to note that when $\alpha = 1$, the fBm simplifies to the Wiener process, resulting in the number of mutations in a lineage during a specific time period t being described as a Brownian motion.

3 Data-driven model validation

Viruses have frequently served as a valuable model system for investigating evolution because of their high mutability and rapid evolutionary changes [14]. In this study, SARS-CoV-2 viral DNA sequences were employed to validate the proposed model due to the comprehensive coverage of the virus's evolution and the availability of high-quality data.

For each sequence in the dataset (all available viral sequences collected in the United Kingdom up to May 2022), the number of mutations was computed by comparing it to the reference SARS-CoV-2 genome sequence (NC_045512.2). Next, the number of mutations were binned in a weekly manner, and the mean and variance of the number of mutations were computed for each week and variant of concern (VoC) annotated. In particular, to perform computations for variants, only sequences annotated as variant v were considered. Certainly, the number of sequences in week k, dubbed N_k, obeys $N_k = \sum_{v \in V} N_{v,k} + N_{\emptyset k}$, where V is the set of variants and $N_{\emptyset k}$ denotes the number of sequences that are not linked to any variant of V in the kth week.

Thus, if there are $N_{v,k}$ sequences in the kth week that are linked to variant v, the mean and variance of the number of mutations are given by

$$\mathbb{E}\left[m_{v,k}\right] = \frac{1}{N_{v,k}} \sum_{i=1}^{N_{v,k}} m_{v,k,i} \tag{19}$$

$$\mathbb{V}\left[m_{v,k}\right] = \frac{1}{N_{v,k}} \sum_{i=1}^{N_{v,k}-1} \left(m_{v,k,i} - \mathbb{E}\left[m_{v,k}\right]\right)^2 \tag{20}$$

The variances were then fitted following the expression

$$\log\left[\left(\mathbb{V}\left[m_{v,k}\right] - \sigma_0^2\right)/\kappa\right] = \alpha \log k \tag{21}$$

Resulting in subdiffusion in the Primal, Alpha and Omicron BA.1 variants, while weak superdiffusion in the case of the Delta variant (Pearson's correlations in log scale, $P < 10^{-4}$ for Primal, Alpha, and Delta and $P = 0.020$ for Omicron BA.1), which resulted in a significant improvement with respect to the null model (Brownian motion). A more elaborated discussion regarding the biological significance of these results, including the implications of the diffusion exponent α and graphical representation of the fitted parameters, can be found in [15].

4 Closing remarks

Anomalous diffusion is gaining traction as a model for describing a great variety of naturally ocurring processes, starting with the diffusion of microscopic particles in a fluid. Models based on anomalous diffusion patterns may be suited to describe the evolution of living entities, including viruses, as they can account for the long-term dependence of the process, which is not possible with the Brownian motion model.

In this work, we have proposed a novel molecular clock model based on anomalous diffusion, which can be used to describe the number of mutations in a lineage during a specific time period t as a stochastic process with a mean and variance equal to κt and κt^α, respectively, where κ represents the rate of substitutions for a given unit of time and α is the diffusion exponent, which determines the degree of long-term dependence of the process. This model has

been validated using SARS-CoV-2 viral DNA sequences, resulting in a significant improvement with respect to the null model (Brownian motion).

Acknowledgements

We are indebted to numerous scientists and health professionals that contributed to generate a public database of SARS-CoV-2 sequences. We thank Roser Montagud-Martínez and Nicolás Firbas for their useful discussions regarding evolution and phylogenetic inference. This work was supported by the CSIC PTI Global Health (grant SGL2021-03-040 to GR) through the NextGenerationEU Fund (regulation 2020/2094) and the Valencia Regional Government (grant GVA-COVID19/2021/100 to JAC and grant GVA-COVID19/2021/036 to GR). LG was supported by a predoctoral fellowship from the Valencia Regional Government (ACIF/2021/183).

References

1. Zuckerkandl, E. & Pauling, L. Evolutionary divergence and convergence in proteins. In: Bryson, V. & Vogel, H.J., Eds., *Evolving Genes and Proteins*. Academic Press, New York, pp. 97-166 (1965)
2. Kimura, M. Evolutionary rate at the molecular level. *Nature*. **217** pp. 624-626 (1968)
3. Kimura, M., & Ohta, T. On the rate of molecular evolution. *J. Mol. Evol.*. **1** pp. 1-17 (1971)
4. Kimura, M. Molecular evolutionary clock and the neutral theory. *J. Mol. Evol.*. **26** pp. 24-33 (1987)
5. Ho, S. The molecular clock and estimating species divergence. *Nature Education*. **1** pp. 168 (2008)
6. Ho, S. Y. & Duchêne, S. Molecular-clock methods for estimating evolutionary rates and Timescales. *Mol. Ecol.*. **23** pp. 5947-5965 (2014)
7. Park, S. Y., Love, T. M., Perelson, A. S., Mack, W. J. &Lee, H. Y. Molecular clock of HIV-1 envelope genes under early immune selection. *Retrovirology*. **13** pp. 38 (2016)
8. Jenkins, G., Rambaut, A., Pybus, O. & Holmes, E. Rates of molecular evolution in RNA viruses: a quantitative phylogenetic analysis. *J. Mol. Evol.*. **54** pp. 156-165 (2002)
9. Langley, C. H., & Fitch, W. M. An examination of the constancy of the rate of molecular evolution. *J. Mol. Evol.*. **3** pp. 161-177 (1974)
10. Bedford, T., Wapinski, I. & Hartl, D. Overdispersion of the molecular clock varies between yeast, Drosophila and mammals. *Genetics*. **179** pp. 977-984 (2008)
11. Ayala, F. Molecular clock mirages. *BioEssays*. **21** pp. 71-75 (1999)
12. Kumar, S. Molecular clocks: four decades of evolution. *Nat. Rev. Genet.*. **6** pp. 654-662 (2005)
13. Muñoz-Gil, G., Volpe, G., Garcia-March, M.A. et al. Objective comparison of methods to decode anomalous diffusion. *Nat. Commun.*. **12** 6253 (2021).
14. Koonin, E. & Dolja, V. A virocentric perspective on the evolution of life. *Curr. Opin. Virol.*. **3** pp. 536-557 (2013)
15. Goiriz, L., Ruiz, R., Garibo-i-Orts, Ò., Conejero, J. A., & Rodrigo, G. A variant-dependent molecular clock with anomalous diffusion models SARS-COV-2 evolution in humans. *Proc. Natl. Acad. Sci. USA*. **120** pp. e2303578120 (2023)