УДК 575.224.4

A. O. Trufanova, V. A. Yurov

THE EQUILIBRIUM STATES OF A mtDNA MUTAGENESIS: AN ANALYTICS APPROACH

Поступила в редакцию 20.03.2022 г.
Рецензия от 03.04.2022 г.

The article is dedicated to the task of construction and analysis of exactly solvable mathematical model for the mutagenesis in mitochondrial DNA (mtDNA). We begin by demonstrating that the average amounts of all four types of nucleotides (adenine A, guanine G, cytosine C and thymine T) is determined by a system of four O.D.E.'s of first order, and that this system is further reducible to a single linear inhomogeneous O.D.E. of third order. Next, we classify all possible solutions to that equation and derive an explicit general form of equilibrium (i.e. asymptotic) states of mtDNA as a function of a mitochondrial mutational spectrum.

Работа посвящена точному математическому моделированию мутагенеза в митохондриальном ДНК (мтДНК). Показано, что динамика среднего количества четырех нуклеотидов (аденина A, гуанина G, цитозина C и тимина T) в мтДНК описывается системой из четырех обыкновенных дифференциальных уравнений первого порядка, которые могут быть сведены к одному линейному неоднородному дифференциальному уравнению третьего порядка. Классифицированы все возможные типы решений этого уравнения и выведен явный вид равновесных (асимптотических) состояний мтДНК как функции митохондриального мутационного спектра.

Keywords: mtDNA, mutagenesis, mutational spectrum, O.D.E., equilibrium states of DNA

Ключевые слова: мтДНК, мутагенез, мутационный спектр, обыкновенные дифференциальные уравнения, равновесные состояния ДНК

1. About 2.5 billion years ago a momentous event has taken place: a free-living α -proteobaterium has been ingested by a different cell and instead of being digested became a symbiont for its host [1; 2]. This seemingly innocuous affair has turned out to be extremely profitable for everyone involved, eventually producing a new very successful lineage of cells that we now call the *eukaryotes*, which now includes such magnificent offsprings as plants, fungi and animals. The mitochondria became a sort of a power plant, producing copious amounts of ATP molecules, that is used as an energy source in various metabolic processes inside of a eukaryotic cell [3, p. 547, 556]. But mitochondria play a number of other important roles in various aspects of eukaryotic life cycles, such as initiating an apoptosis [4], storing the calcium ions [5] and so on. Still, the mitochondria, being a

symbiont, has retained its own mitochondrial DNA (which is usually shortened to mtDNA), which exists and replicated independently of a nuclear DNA. And the mtDNA is interesting in a number of ways. First of all, it is circular, making it quite distinct from the linear nuclear DNA of animals. Second, during the replication of mtDNA its two strands behave very differently, as one of them (so called "heavy strand") spends tens of minutes (30 min to 1 hour) to grow a new complimentary chain — the other one (the "light strand") receiving its new chain almost immediately (see Fig.) [6]. This disparity means that if any kind of damage occurs during the mtDNA replication (say, due to such reactive oxygen species as peroxides or hydroxyl radicals) — it will most certainly occur at a heavy strand, as the one that spends more time exposed to the elements. In fact, as has been demonstrated in [7], it is this endogenous mutational mechanism that is responsible for a majority of mutations in the human mtDNA.

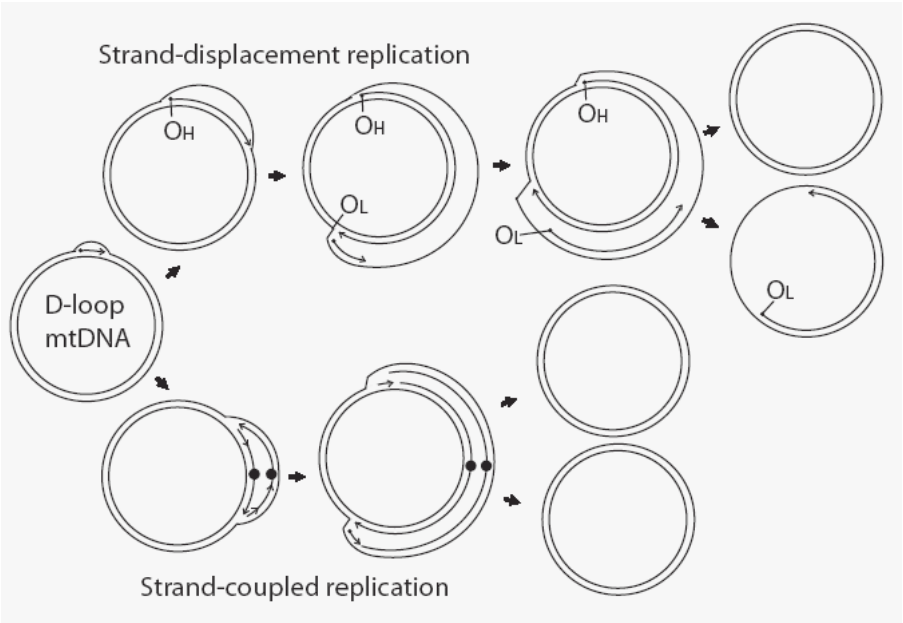

Fig. Two proposed mechanisms of replication of mtDNA.
The evidences suggest that the strand-displacement (the upper one)
is the correct one. (Adapted from [6])

Hence, we end up with a very interesting observation. Suppose we want to study the mutagenesis of mtDNA — for example, to use it as an indicator for a cellular aging [8]. Taking into account the aforementioned mtDNA strands asymmetry, we can safely limit ourselves to describing the mutagenesis on heavy strand only. This in turn implies that instead of sticking to the classical Kimura model [9; 10], originally designed for the nuclear DNA and making no distinction between the mutagenesis of individual nucleotides, we can now attempt to construct a mathematical model that accounts for every type of nucleotide on the heavy strand *separately*.

So, how shall we can study the process and the final outcome of neutral mutagenesis of genomes characterized by strongly asymmetric replication (not limited to mtDNA, but also including the COVID-19 RNA etc.). The main idea would be to employ an entirely **analytical model**, based in the ordinary differential equations.

For that end, let us consider a sample consisting of a neutral part of mtDNA (either in its entirety or a part of it), and denote the total number of the nucleotides on it by N. Next, let us denote by $\boldsymbol{a}, \boldsymbol{g}, \boldsymbol{\tau}, \boldsymbol{c}$ the average amounts of A, G, T and C nucleotides on our sample[1]. All of these functions depend on time t.

Next, we introduce the *mutation spectrum*: the twelve probabilities of a given nucleotide (say, A) to mutate to any other (i.e. to G, T or C). We will denote these probabilities by r_{xy}, where x denotes the type of an original nucleotide and y — the new nucleotide (e.g., r_{ag} is a probability of an $A > G$ mutation). The rate of any type of mutation (say, $G > A$) is proportional to: a corresponding probability r_{ga}, to a time interval Δt (the more time has passed the more probable it will be for *G* to turn to *A*), and, finally, to a total amount of original nucleotide (function $\boldsymbol{g}(t)$).

Say, we want to look at the time-dependent evolution of the A nucleotides. If at time t we have an average of $a(t)$ adenine on our sample, how much will remain after $\Delta t \ll 1$ has passed?

In order to answer this question we have to account for both a surplus of brand new A nucleotides, produced from G, T and C; as well as for the removal of at least some old A nucleotides by means of mutation.

This leads to the following relation:

$$a(\mathrm{t}+\Delta t) \approx a(t) + \Delta t \cdot \left[r_{ga} \cdot g(t) + r_{\tau a} \cdot \tau(t) + r_{ca} \cdot c(t) - \left(r_{ag} + r_{a\tau} + r_{ac} \right) \cdot a(t) \right]$$

$$\frac{a(\mathrm{t}+\Delta t) - a(t)}{\Delta t} \approx r_{ga} \cdot g(t) + r_{\tau a} \cdot \tau(t) + r_{ca} \cdot c(t) - \left(r_{ag} + r_{a\tau} + r_{ac} \right) \cdot a(t)$$

Naturally, if we then take a limit $\Delta t \to 0$, then the approximate equality will be exact (if have averaged over sufficiently many RNA's!) and we'll end with an ordinary linear homogeneous differential equation of 1st order. Repeating the same process for the remaining 3 nucleotides we'll end up with the following closed system of O.D.E.'s:

$$\frac{da}{dt} = -\left(r_{ag} + r_{a\tau} + r_{ac} \right) \cdot a(t) + r_{ga} \cdot g(t) + r_{\tau a} \cdot \tau(t) + r_{ca} \cdot c(t), \qquad (1)$$

$$\frac{dg}{dt} = r_{ag} \cdot a(t) - \left(r_{ga} + r_{g\tau} + r_{gc} \right) \cdot g(t) + r_{\tau g} \cdot \tau(t) + r_{cg} \cdot c(t), \qquad (2)$$

[1] Naturally, in order for our model to work the average has to be taken over sufficiently many copies of COVID-19's RNA — otherwise the functions $\boldsymbol{a}, \boldsymbol{g}, \boldsymbol{\tau}, \boldsymbol{c}$ will be neither continuous nor differentiable.

$$\frac{d\tau}{dt} = r_{a\tau}\cdot a(t) + r_{g\tau}\cdot g(t) - \left(r_{\tau a} + r_{\tau g} + r_{\tau c}\right)\cdot \tau(t) + r_{c\tau}\cdot c(t), \qquad (3)$$

$$\frac{dc}{dt} = r_{ac}\cdot a(t) + r_{gc}\cdot g(t) + r_{\tau c}\cdot \tau(t) - \left(r_{ca} + r_{cg} + r_{c\tau}\right)\cdot c(t). \qquad (4)$$

There are a couple of approaches one can utilize to solve the system (1)—(4). For example, one can rewrite this system as a single matrix equation of first order:

$$\frac{d}{dt}\begin{pmatrix} a \\ g \\ \tau \\ c \end{pmatrix} = M\cdot\begin{pmatrix} a \\ g \\ \tau \\ c \end{pmatrix}$$

and then look for the eigenvalues of the 4x4 matrix M (see, for example, the brute-forcing approach provided in [11]). However, if we are going to study all the possible modes of mutagenesis, it will be worthwhile to solve the system (1)—(4) in general, since it will only provide us will all permissible solutions for (1)—(4), but can also be used as an approach in the more general case where the mutational spectrum is a function of time.

With that said, let us get back to our system. Note, that if we sum up these four equation, we'll have

$$\frac{da}{dt} + \frac{dg}{dt} + \frac{d\tau}{dt} + \frac{dc}{dt} = 0 \qquad \Rightarrow \qquad a + g + \tau + c = N = const, \qquad (5)$$

which is both consistent with our expectations, and allows us to reduce the system (1—4) to just 3 equations by simply removing one of the functions using (5). Say, we get rid of $c(t)$. Then

$$\begin{cases} \dfrac{da}{dt} = k_{11}\cdot a(t) + k_{12}\cdot g(t) + k_{13}\cdot \tau(t) + f_1 \\ \dfrac{dg}{dt} = k_{21}\cdot a(t) + k_{22}\cdot g(t) + k_{23}\cdot \tau(t) + f_2 \\ \dfrac{d\tau}{dt} = k_{31}\cdot a(t) + k_{32}\cdot g(t) + k_{33}\cdot \tau(t) + f_3 \end{cases}, \qquad (6)$$

where $f_1 = r_{ca}\cdot N$, $f_2 = r_{cg}\cdot N$, $f_3 = r_{c\tau}\cdot N$, and k_{ij} are the elements of matrix K:

$$K = \begin{pmatrix} -\left(r_{ag} + r_{a\tau} + r_{ac} + r_{ca}\right) & r_{ga} - r_{ca} & r_{\tau a} - r_{ca} \\ r_{ag} - r_{cg} & -\left(r_{ga} + r_{g\tau} + r_{gc} + r_{cg}\right) & r_{\tau g} - r_{cg} \\ r_{a\tau} - r_{c\tau} & r_{g\tau} - r_{c\tau} & -\left(r_{\tau a} + r_{\tau g} + r_{\tau c} + r_{c\tau}\right) \end{pmatrix}. \qquad (7)$$

At this juncture it is useful to make an additional reasonable assumption: that the mutation spectrum is time-independent (i.e. that all rates of mutation are constant). If it is true then it will be possible to reduce the resulting

system even further — up to a single equation! For example, it is possible to show that $a(t)$ will satisfy the following *inhomogeneous linear differential equation of* ***third order***:

$$a''' + r \cdot a'' + d \cdot a' + \Delta \cdot a = f, \tag{8}$$

where r, d, Δ, $f > 0$ are all positive (!) constants that are derived from k_{ij} as follows:

$$r = -\left(k_{11} + k_{22} + k_{33}\right), \tag{9}$$

$$d = \begin{vmatrix} k_{11} & k_{12} \\ k_{21} & k_{22} \end{vmatrix} + \begin{vmatrix} k_{11} & k_{13} \\ k_{31} & k_{33} \end{vmatrix} + \begin{vmatrix} k_{22} & k_{23} \\ k_{32} & k_{33} \end{vmatrix}, \tag{10}$$

$$\Delta = -\begin{vmatrix} k_{11} & k_{12} & k_{13} \\ k_{21} & k_{22} & k_{23} \\ k_{31} & k_{32} & k_{33} \end{vmatrix} = -\det K, \tag{11}$$

$$f = \begin{vmatrix} f_1 & k_{12} & k_{13} \\ f_2 & k_{22} & k_{23} \\ f_3 & k_{32} & k_{33} \end{vmatrix}. \tag{12}$$

Interestingly, the multiple r has a very simple form if we revert back from k_{ij} to r_{xy}: it is just a sum of all 12 probabilities r_{xy}. Unfortunately, the remaining 3 multiples are less fortunate, and are actually easier to handle written in the forms of determinant (10—12).

It is easy to show that the general solution of our third-order equation (8) has a form (see, for example, [12; 13]):

$$a(t) = \tilde{a}(t) + \frac{f}{\Delta}, \tag{13}$$

where $\tilde{a}(t)$ is a general solution of the following linear homogeneous equation:

$$a''' + r \cdot a'' + d \cdot a' + \Delta \cdot a = 0. \tag{14}$$

The exact form of $\tilde{a}$ will depend on the roots of the companion *characteristic equation* [12; 13]:

$$\lambda^3 + r \cdot \lambda^2 + d \cdot \lambda + \Delta = 0. \tag{15}$$

There exist four possibilities (here C_j denotes arbitrary constants):

1. There are 3 distinct **negative** real roots $\lambda_1, \lambda_2, \lambda_3 < 0$. Then

$$\tilde{a} = C_1 \cdot e^{\lambda_1 \cdot t} + C_2 \cdot e^{\lambda_2 \cdot t} + C_3 \cdot e^{\lambda_3 \cdot t} \to 0 \text{ when } t \to +\infty.$$

2. There are 3 **negative** real roots $\lambda_1, \lambda_2, \lambda_3 < 0$, but $\lambda_2 = \lambda_3$. Then

$$\tilde{a} = C_1 \cdot e^{\lambda_1 \cdot t} + (C_2 \cdot t + C_3) \cdot e^{\lambda_2 \cdot t} \to 0 \text{ when } t \to +\infty.$$

3. There are three identical **negative** real roots $\lambda_1 = \lambda_2 = \lambda_3 = -\frac{r}{3} < 0$:

$$\tilde{a} = (C_1 \cdot t^2 + C_2 \cdot t + C_3) \cdot e^{-r \cdot t/3} \to 0 \text{ when } t \to +\infty.$$

4. There is one real **negative** root $\lambda_1 < 0$ and two complex conjugate roots $\lambda_2 = \Lambda + i\omega$, $\lambda_3 = \Lambda - i\omega$, where $\Lambda < 0$. In this case

$$\tilde{a} = C_1 \cdot e^{\lambda_1 \cdot t} + (C_2 \cdot \cos \omega t + C_3 \cdot \sin \omega t) \cdot e^{\Lambda \cdot t} \to 0 \text{ when } t \to +\infty.$$

Note that in all four cases the overall behavior of the solution of (14) is controlled by asymptotic convergence to zero for sufficiently large $t > 0$. Thus, given sufficiently large time, the average amount of *A* nucleotides will ***always*** converge to the *equilibrium state*:

$$\lim_{t \to +\infty} a(t) = \frac{f}{\Delta}, \tag{16}$$

where

$$f = N \cdot \begin{vmatrix} r_{ca} & r_{ga} - r_{ca} & r_{\tau a} - r_{ca} \\ r_{cg} & -\left(r_{ga} + r_{g\tau} + r_{gc} + r_{cg}\right) & r_{\tau g} - r_{cg} \\ r_{c\tau} & r_{g\tau} - r_{c\tau} & -\left(r_{\tau a} + r_{\tau g} + r_{\tau c} + r_{c\tau}\right) \end{vmatrix}, \tag{17}$$

$$\Delta = \begin{vmatrix} r_{ag} + r_{a\tau} + r_{ac} + r_{ca} & r_{ca} - r_{ga} & r_{ca} - r_{\tau a} \\ r_{cg} - r_{ag} & r_{ga} + r_{g\tau} + r_{gc} + r_{cg} & r_{cg} - r_{\tau g} \\ r_{c\tau} - r_{a\tau} & r_{c\tau} - r_{g\tau} & r_{\tau a} + r_{\tau g} + r_{\tau c} + r_{c\tau} \end{vmatrix}, \tag{18}$$

which depend only on the total amount N of nucleotides in the sample and the mutational spectrum. We would like to emphasize that the resulting analytical formulas do not in any way depend on either the initial state of the sample (i.e. on a_0, g_0, c_0, τ_0) or on the particulars of how the equilibrium is reached — all of that is effectively swept under a rug by the asymptotic behavior of exponential functions in the solutions for $\tilde{a}$.

We conclude by noting that the similar approach can also be used for the remaining three nucleotides. But the formulas for their equilibrium states might also be very easily derived directly from (16—18). All one has to do would be to take these formulas and simply switch the indices. For example, the equilibrium state of *T* nucleotide can be gained by switching $a \leftrightarrow \tau$ in (17—18), so that r_{ca} becomes $r_{c\tau}$, whereas $r_{c\tau}$ is replaced by r_{ca}, etc.

Список литературы

1. *Gray M.W., Burger G., Lang B.F.* Mitochondrial evolution // Science. 1999. Vol. 283. P. 1476—1481.

2. *Fan L., Wu D., Goremykin V., Xiao J.* Phylogenetic analyses with systematic taxon sampling show that mitochondria branch within Alphaproteobacteria // Nature Ecology & Evolution. 2020. Vol. 4 (9). P. 1213—1219.

3. *Voet D., Voet J.G., Pratt C.W.* Fundamentals of Biochemistry. 2nd ed. John Wiley and Sons, Inc, 2006.

4. *Wang C., Youle R.J.* The Role of Mitochondria in Apoptosis // Ann. Rev. Genet. 2009. Vol. 43. P. 95—118.

5. *Santulli G., Xie W., Reiken S.R., Marks A.R.* Mitochondrial calcium overload is a key determinant in heart failure // The Proceedings of the National Academy of Sciences. Vol. 112 (36). P. 11389—11394.

6. *Brown T.A., Cecconi C., Tkachuk A.N. et al.* Replication of mitochondrial DNA occurs by strand displacement with alternative light-strand origins, not via a strand-coupled mechanism // Genes. Dev. 2005. Vol. 19 (20). P. 2466—2476.

7. *Ju Y.S., Alexandrov L.B., Gerstung M. et al.* Origins and functional consequences of somatic mitochondrial DNA mutations in human cancer // Elife. 2014. 2014. Oct. 1; 3: e02935. doi: 10.7554/eLife.02935.

8. *Mikhaylova A.G., Mikhailova A.A., Ushakova K. et al.* Mammalian mitochondrial mutational spectrum as a hallmark of cellular and organismal aging // bioRxiv:/ 10.1101/589168v3. doi: https://doi.org/10.1101/589168.

9. *Kimura M.* Evolutionary Rate at the Molecular Level // Nature. 1968. Vol. 217. P. 624—626.

10. *Kimura M.* The neutral theory of molecular evolution. Cambridge University Press, 1984.

11. *Novembre J.* Computing Stationary Distributions of a Discrete Markov Chain. URL: https://stephens999.github.io/fiveMinuteStats/markov_chains_discrete_stationary_dist.html (дата обращения: 30.02.2022).

12. *Тихонов А.Н., Васильева А.Б., Свешников А.Г.* Дифференциальные уравнения. М., 1980.

13. *Зенков А.В.* Дифференциальные уравнения, несобственные интегралы и интегралы, зависящие от параметра. Екатеринбург, 2010.

The authors

Arina O. Trufanova, Master's Student, Immanuel Kant Baltic Federal University, Russia.

E-mail: arishka_39@mail.ru

Dr Valerian A. Yurov, Associate Professor, Immanuel Kant Baltic Federal University, Russia.

E-mail: vayt37@gmail.ru

Об авторах

Арина Олеговна Труфанова — магистрант, Балтийский федеральный университет им. И. Канта, Россия.

E-mail: arishka_39@mail.ru

Валериан Артемович Юров — канд. физ.-мат. наук, доц., Балтийский федеральный университет им. И. Канта, Россия.

E-mail: vayt37@gmail.ru