genes

MDPI

Article

Note on DNA Analysis and Redesigning Using Markov Chain

Maciej Zakarczemny *,† and Małgorzata Zajęcka †

Department of Applied Mathematics, Faculty of Computer Science and Telecommunications, Cracow University of Technology (CUT), 24 Warszawska Street, 31-155 Cracow, Poland; malgorzata.zajecka@pk.edu.pl
* Correspondence: mzakarczemny@pk.edu.pl
† These authors contributed equally to this work.

Abstract: The paper contains a discussion on mathematical modifying and redesigning DNA with the use of Markov chains. We give a simple mathematical technique for overwriting missing parts of DNA. With a certain probability (without even knowing the function of the missing codon) we can find a synonymous codon, so that there is no frequency change in amino acid sequences of proteins. We use Markov Chain to analyze the dependencies in DNA sequence of the human gene Alpha 1,3-Galactosyltransfe rase 2. We include a theoretical introduction which facilitates the understanding of the paper for non-mathematicians, especially for biologists not familiar with the theory of Markov chains.

Keywords: DNA; Markov chain; human gene

MSC: 92B05; 60J20

Citation: Zakarczemny, M.; Zajęcka, M. Note on DNA Analysis and Redesigning Using Markov Chain. *Genes* **2022**, *13*, 554. https://doi.org/10.3390/genes13030554

Academic Editor: Stefano Lonardi

Received: 28 January 2022
Accepted: 9 March 2022
Published: 21 March 2022

1. Introduction

1.1. Motivation and Methods

Modeling and analyzing DNA sequences using statistical methods have been a challenge for statisticians and biologists for many years. For many years, the most common approach was based on the theory of Markov chains. Well known simple models appeared in the mid-1980s in the papers by B.E. Blaisdell [1] and V. Brendel, J.S. Beckmann, E.N. Trifonov [2]. This was later followed by more advanced models developed to study different biological aspects of DNA (see [3–5]) that have been also using Markov chains (both homogeneous and non-homogeneous ones, possibly of higher orders). There are some statistical results showing that first-order Markov chains are not an adequate model for DNA sequences (see, e.g., [6]). Moreover, the latest comparison studies (see [7]) show that in general DNA sequencing by models based on even higher order Markov chains does not fit perfectly. Thus the latest research in DNA sequencing in bioinformatics is focusing on deep learning methods (see [8]). The methods presented in this paper, however, are local rather than statistical. We apply our model to DNA sequences less than 55 base pairs long, which is not enough for statistical methods. Markov chain theory is being applied to modeling music and literature. For example, a random song generated by a Markov chain based on some given piece of music, can achieve a similarity level comparable to this piece (see [9–11]). The question arises whether DNA can be handled similarly to a piece of music. Numerous attempts to write an understandable text as the realization of a low-order Markov chain have not proved successful. For high-order Markov chains, however, such a text becomes understandable as it contains complete sentences from the original text (see [12–14]). Therefore, the question whether Markov chains of any order are suitable to study DNA sequencing becomes the question about the complexity of the DNA structure. In this paper we show methods for filling in short gaps in DNA sequences. The results obtained by our method are then compared with the original DNA sequence.

1.2. Notations

We consider a probability space $(\Omega, \mathcal{F}, P)$, where $\Omega = \{S_1, S_2, \ldots, S_n\}$ denotes a finite state space, i.e., the set of all possible results of an experiment, $\mathcal{F}$ is a σ-field of events, and $P : \mathcal{F} \to [0, 1]$ denotes a probabilistic measure. By *random variable* we denote a function $Z : \Omega \to \mathbb{R}$ such that for any $a \in \mathbb{R}$ a preimage $Z^{-1}((-\infty, a])$ is in the σ-field $\mathcal{F}$. In our case a random variable will allow us to assign natural numbers to states from the state space Ω: let E be a single experiment with possible results forming Ω, if in t-th repetition (i.e., in time t) of experiment E we obtain result $S_j \in \Omega$, then we put $Z_t = j$, where Z_t is a random variable with natural values.

1.3. Stochastic Matrices

Definition 1. *A matrix $\Pi = (p_{ij})_{i,j\in\{1,\ldots,n\}}$ is called stochastic if all p_{ij} are non-negative and for any i we have $\sum_{j=1}^{n} p_{ij} = 1$ (right stochastic matrix) or for any j we have $\sum_{i=1}^{n} p_{ij} = 1$ (left stochastic matrix). A doubly stochastic matrix is both left stochastic and right stochastic. A vector with non-negative real elements is a stochastic vector if its elements sum to one.*

Remark 1. *Rows of right (columns of left) stochastic matrix are row (vertical) stochastic vectors.*

Using basic algebra one can prove the following remark.

Remark 2. *The product of two right (two left) stochastic matrices is a right (left) stochastic matrix.*

Definition 2. *A square matrix $A = (a_{ij})_{i,j\in\{1,\ldots,n\}}$ is called irreducible if for any partition $S \cup T = \{1, \ldots, n\}$, $S \cap T = \varnothing$ there exists $s \in S$, $t \in T$ such that $a_{st} \neq 0$.*

Definition 3. *Let A be a square matrix. An eigenvalue of A is a complex number λ such that* $\det(A - \lambda\mathbf{I}) = 0$ *(i.e., λ is a zero with multiplicity $k \geq 1$ of the* characteristic polynomial *$p_A(\lambda) = \det(A - \lambda\mathbf{I})$), where $\mathbf{I}$ denotes the identity matrix. The set of all eigenvalues is called spectrum.*

Definition 4. *The eigenvector of a of a square matrix A is a column vector v such that $Av = \lambda v$, where λ is an eigenvalue of A. The left eigenvector of a square matrix A is a column vector w such that $w^T A = \lambda w^T$, where λ is an eigenvalue of A. Some authors define the left eigenvector as a row vector w^T.*

2. Markov Chains

A Markov chain is a sequence of random variables forming a probabilistic model describing a memoryless type of dependency: the future may depend only on the present and must be independent of the past.

2.1. Model

1. E is an experiment with possible results forming a finite set $\Omega = \{S_1, S_2, \ldots, S_n\}$;
2. $\mathcal{S} = \{1, 2, \ldots, n\}$ is a state space associated with Ω;
3. $(\Omega, \mathcal{F}, P)$ is a probability space, where $\mathcal{F} \subset 2^{\Omega}$ is a σ-field and P is a probabilistic measure $P : \mathcal{F} \to [0, \infty)$;
4. Z_t is a random variable defined as follows: if in t-th repetition of experiment E we obtain result $S_j \in \Omega$, then we put $Z_t = j \in \mathcal{S}$.

Definition 5. *A sequence of random variables $(Z_t)_{t=0}^{\infty}$ with values in a state space $\mathcal{S}$ is a* Markov chain *if for all $t \in \mathbb{N}$ and all $j_0, j_1, \ldots, j_t \in \mathcal{S}$*

$$P(Z_t = j_t | Z_0 = j_0, Z_1 = j_1, \ldots, Z_{t-1} = j_{t-1}) = P(Z_t = j_t | Z_{t-1} = j_{t-1}) \quad (1)$$

if only

$$P(Z_0 = j_0, Z_1 = j_1, \ldots, Z_{t-1} = j_{t-1}) > 0.$$

Remark 3. *In the general case a state space $\mathcal{S}$ can be an arbitrary countable subset of $\mathbb{N}$ (positive integers).*

Remark 4. *The Equation (1) is called Markov property and its right-hand side is called a transition operator i.e., the probability of moving from state j_{t-1} to state j_t in one step.*

Definition 6. *Let $(Z_t)_{t=0}^{\infty}$ be a Markov chain. For $t \geq 1$ we call a stochastic matrix $\Pi(t) = (p_{ij}(t))_{i,j\in\mathcal{S}}$ a transition Matrix of a Markov chain (Z_t) in time t if*

$$p_{ij}(t) = P(Z_t = j | Z_{t-1} = i) \tag{2}$$

for all i such that $P(Z_{t-1} = i) > 0$. Since the total of transition probability from one state to all other states must be equal to one, thus this matrix is a right stochastic matrix.

Definition 7. *Let $(Z_t)_{t=0}^{\infty}$ be a Markov chain. If $P(Z_t = j_t | Z_{t-1} = j_{t-1})$ is independent of t, then we call this Markov chain homogeneous.*

Remark 5. *For our convenience we index states by t, but we remind the reader of the following property of homogeneous Markov chains*

$$P(Z_t = j_t | Z_{t-1} = j_{t-1}) = P(Z_{t+m} = j_t | Z_{t+m-1} = j_{t-1}),\ t, m \in \mathbb{N}.$$

Remark 6. *If a Markov chain $(Z_t)_{t=0}^{\infty}$ is homogeneous then there exists a stochastic matrix $\Pi = (p_{ij})_{i,j\in\mathcal{S}}$ such that for all $t \geq 1$ transition matrix $\Pi(t) = \Pi$, where each value p_{ij} is the probability of moving from state i to state j in one step.*

Now we are going to consider probabilities of moving from state to state in larger number of steps.

Definition 8. *Let $(Z_t)_{t=0}^{\infty}$ be a homogeneous Markov chain. For $m \geq 1$ we call a stochastic matrix $\Pi^{(m)} = (p_{ij}^{(m)})_{i,j\in\mathcal{S}}$ a transition Matrix of a Markov chain (Z_t) in m steps if*

$$p_{ij}^{(m)} = P(Z_m = j | Z_0 = i) \tag{3}$$

for all i such that $P(Z_0 = i) > 0$.

Theorem 1. *Let $(Z_t)_{t=0}^{\infty}$ be a homogeneous Markov chain. Then $\Pi^{(m)} = \Pi^m$.*

Proof. For $m = 1$ theorem is true because $\Pi^{(1)} = \Pi$. For $m \geq 2$ from the law of total probability we obtain

$$p_{ij}^{(m)} = P(Z_m = j | Z_0 = i)$$
$$= \sum_{s\in\mathcal{S}} P(Z_{m-1} = s | Z_0 = i) P(Z_m = j | Z_{m-1} = s) = \sum_{s\in\mathcal{S}} p_{is}^{(m-1)} p_{sj}.$$

Hence $\Pi^{(m)} = \Pi^{(m-1)}\Pi$, thus $\Pi^{(m)} = \Pi^m$. □

Example 1 (Random walk on a complex plane)**.** *Identify adenine with $-i$, cytosine with 1, guanine with -1, and thymine with i. Let $(U_n)_{n=0}^{\infty}$ be a sequence of independent random variables such that for any $n \geq 1$:*

$$P(U_n = 1) = p_1,\ P(U_n = i) = p_2,\ P(U_n = -1) = p_3,\ P(U_n = -i) = p_4,$$

where $p_1 + p_2 + p_3 + p_4 = 1$, *and* $P(U_0 = 0) = 1$.

For $n \geq 1$ *define* $Z_n = U_0 + U_1 + \ldots + U_n$ *which is a random variable with values in state space of Gaussian integers* $\mathbb{Z}[i] = \{a + bi : a, b \in \mathbb{Z}\}$. *Gaussian integers form a commutative ring and 2-dimensional integer lattice. We show that* $(Z_n)_{n=0}^{\infty}$ *is a Markov sequence.*

Let $j_0, j_1, \ldots, j_n \in \mathbb{Z}[i]$ *be such that* $P(Z_0 = j_0, Z_1 = j_1, \ldots, Z_{n-1} = j_{n-1}) > 0$. *Then, from definition of* Z_n,

$$\begin{aligned} &P(Z_n = j_n | Z_0 = j_0, Z_1 = j_1, \ldots, Z_{n-1} = j_{n-1}) \\ &= \frac{P(Z_0 = j_0, Z_1 = j_1, \ldots, Z_{n-1} = j_{n-1}, Z_n = j_n)}{P(Z_0 = j_0, Z_1 = j_1, \ldots, Z_{n-1} = j_{n-1})} \\ &= \frac{P(U_0 = j_0, U_1 = j_1 - j_0, \ldots, U_{n-1} = j_{n-1} - j_{n-2}, U_n = j_n - j_{n-1})}{P(U_0 = j_0, U_1 = j_1 - j_0, \ldots, U_{n-1} = j_{n-1} - j_{n-2})} \\ &= \frac{P(U_0 = j_0)P(U_1 = j_1 - j_0) \cdot \ldots \cdot P(U_{n-1} = j_{n-1} - j_{n-2})P(U_n = j_n - j_{n-1})}{P(U_0 = j_0)P(U_1 = j_1 - j_0) \cdot \ldots \cdot P(U_{n-1} = j_{n-1} - j_{n-2})} \\ &= P(U_n = j_n - j_{n-1}) \end{aligned}$$

because variables $(U_n)_{n=0}^{\infty}$ *were independent. Observe that variables* U_n *and* Z_{n-1} *are also independent, thus*

$$\begin{aligned} P(U_n = j_n - j_{n-1}) &= \frac{P(U_n = j_n - j_{n-1})P(Z_{n-1} = j_{n-1})}{P(Z_{n-1} = j_{n-1})} \\ = \frac{P(Z_{n-1} = j_{n-1}, U_n = j_n - j_{n-1})}{P(Z_{n-1} = j_{n-1})} &= \frac{P(Z_{n-1} = j_{n-1}, Z_n = j_n)}{P(Z_{n-1} = j_{n-1})} \\ &= P(Z_n = j_n | Z_{n-1} = j_{n-1}). \end{aligned}$$

One can show that $E(|Z_n|)$*—the expected translation distance after n steps is of order* $\sqrt{n}$, *more precise,* $\lim_{n\to\infty} \frac{E(|Z_n|)}{\sqrt{n}} = \sqrt{\frac{\pi}{2}}$. *With our identification of adenine* $(-i)$, *cytosine* (1), *guanine* (-1), *and thymine* (i) *using Markov chain* $(Z_n)_{n=0}^{\infty}$ *we can consider probability that from n-th to m-th place in DNA strand,* $n < m$, *number of adenine equals number of thymine and, simultaneously, number of cytosine equals number of guanine. This situation means that our random walk made a loop, that is* $Z_n = Z_m$ *(see Figure 1). If one needs to research other pairwise equalities it suffices to change the identification. Denote* $Z[n_1, n_2] = \{Z_n : n_1 \leq n \leq n_2\}$. *One can show that for all* $k \geq 2$ *we have* $P(Z[0, k] \cap Z[2k, 3k] \neq \varnothing) > 0$. *That means that with positive probability there exists* $n \in [0, k]$ *and* $m \in [2k, 3k]$ *such that* $Z_n = Z_m$, *i.e., we have a loop.*

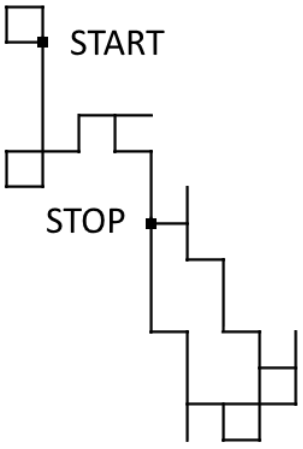

Figure 1. Fifty-four step random walk from a central point on a complex plane. Based on DNA sequence from Example 2.

2.2. Classification of States and Chains

In this subsection, we will give some mathematical background and also state some well known results, see [7,15].

Definition 9. *A state i is called accessible from state j if there exists $n \geq 0$ such that $P(Z_n = i|Z_0 = j) > 0$. If state i is accessible from state j and vice versa we say that states i and j communicate.*

Observe that communication is an equivalence relation that divides states into equivalence classes called *communicating classes*.

Definition 10. *A Markov chain is called irreducible if its state space forms a single communicating class.*

In other words, in irreducible Markov chain it is possible to get from any state to any state (every two states communicate).

Definition 11. *A state i is called inessential if there exists a state j and $n \geq 1$ such that $P(Z_n = j|Z_0 = i) > 0$ and $P(Z_k = i|Z_0 = j) = 0$ for any $k \geq 0$. A state is essential if it is not inessential.*

Denote $f_{ij}^{(n)} = P(Z_n = j, Z_{n-1} \neq j, \ldots, Z_1 \neq j|Z_0 = i)$, $F_{ij} = \sum_{n=1}^{\infty} f_{ij}^{(n)}$, $P_i = \sum_{n=1}^{\infty} p_{ii}^{(n)}$ and define $N_i = \sum_{n=1}^{\infty} \mathbb{1}_{\{Z_n=i\}}$, $M_i = \sum_{n=1}^{\infty} \mathbb{1}_{\{T_{ii}\geq n\}}$, where $T_{ij} = \inf\{n \in \mathbb{N} : Z_n = j\}$ when $Z_0 = i$. Then $f_{ij}^{(n)}$ is a probability that we access state j from state i exactly in n steps, F_{ij} is a probability that we ever access state j from state i, P_i a trace of the transition matrix in n steps, N_i is a random variable that counts how many times state i is accessed and M_i is a random variable that counts how many steps are needed to reappear in state i for the first time.

Remark 7.

$$P_i = \sum_{n=1}^{\infty} P(Z_n = i|Z_0 = i) = \sum_{n=1}^{\infty} \mathcal{E}(\mathbb{1}_{\{Z_n=i\}}|Z_0 = i) = \mathcal{E}(N_i|Z_0 = i),$$

where $\mathcal{E}$ denotes an expected value. Thus P_i is an average time a Markov chain is in state i (averagely how many times Markov chain is in state i). If we define $\mu_i = \sum_{n=1}^{\infty} P(T_{ii} \geq n|Z_0 = i)$ then $\mu_i = \mathcal{E}(M_i|Z_0 = i)$. Thus μ_i is an average number of steps needed to reappear in state i.

Definition 12. *A state i is called recurrent if $F_{ii} = 1$. If $F_{ii} < 0$ then state i is called transient.*

We will need the following result.

Theorem 2 (see [15])**.**

(a) *A state i is recurrent if and only if $P(N_i = \infty|Z_0 = i) = 1$.*
(b) *A state i is transient if and only if $P(N_j < \infty|Z_0 = i) = 1$.*

Theorem 3.

(a) *A state i is transient if and only if $P_i < \infty$.*
(b) *A state i is recurrent if and only if $P_i = \infty$.*

Proof. Note that for all states i, j and natural n we have

$$p_{ij}^{(n)} = \sum_{m=1}^{n} f_{ij}^{(m)} p_{jj}^{(n-m)}. \tag{4}$$

It follows from the fact that accessing state j from state i after n steps means that we access state j for the first time after exactly m steps (for some $m \leq n$) and then after next $n - m$ steps we return to it (perhaps reaching state j a few times on the way). Thus we have

$$\sum_{k=1}^{n} p_{ii}^{(k)} = \sum_{k=1}^{n} \sum_{m=0}^{k-1} f_{ii}^{(k-m)} p_{ii}^{(m)} = \sum_{m=0}^{n-1} p_{ii}^{(m)} \sum_{k=m+1}^{n} f_{ii}^{(k-m)}$$
$$\leq \sum_{m=0}^{n} p_{ii}^{(m)} F_{ii} = F_{ii} + F_{ii} \sum_{m=1}^{n} p_{ii}^{(m)}.$$

Hence we get an inequality

$$(1 - F_{ii}) \sum_{m=1}^{n} p_{ii}^{(m)} \leq F_{ii}. \tag{5}$$

Since n is arbitrary, as n tends to infinity we obtain

$$(1 - F_{ii}) P_i \leq F_{ii}. \tag{6}$$

Assume that state i is transient. Then from (6) we obtain $P_i < \infty$. On the other hand if $P_i < \infty$ then $\mathcal{E}(N_i | Z_0 = i) < \infty$, thus $P(N_i = \infty | Z_0 = i) = 0$. From Theorem 2 state i is transient. If i is recurrent, then we must have $P_i = \infty$ (otherwise see proof of point (a)). Now assume $P_i = \infty$. If $F_{ii} < 1$ then $(1 - F_{ii})P_i$ is unbounded and from (6) we get a contradiction. □

Remark 8. *In irreducible Markov chain either all states are recurrent or all states are transient. Thus we call an irreducible Markov chain recurrent or transient, depending on type of states.*

Remark 9. *One can show that for a finite Markov chain (chain with a finite state space) a state is inessential if and only if it is transient, thus a state is essential if and only if it is recurrent.*

Remark 10. *In the case of gene A3GALT2 each state is essential (because all entries in transition matrix are nonzero), thus from Remark 9 each state is recurrent. This also can be shown using properties of transition matrix: evaluate $p_{ii}^{(n)}$ from the matrix Π^n. Then $P_i = \infty$ for each $i \in \{a, c, g, t\}$. From Theorem 3 we once again obtain that each state is recurrent.*

Definition 13. *A state i is called null-recurrent if $\lim_{n\to\infty} p_{ii}^{(n)} = 0$. A state which is not null-recurrent is called positive recurrent.*

Definition 14. *A state i is called periodic with period d_i if $d_i = \mathrm{GCD}\{n > 0 : p_{ii}^{(n)} > 0\} > 1$ (if for all $n > 0$ we have $p_{ii}^{(n)} = 0$ then we put $d_i = \infty$). If $d_i = 1$ state i is called aperiodic.*

Definition 15. *A state which is aperiodic, recurrent, and positive recurrent is called ergodic.*

Remark 11. *In case of our matrix Π all states are ergodic.*

For irreducible matrices we have the following property.

Theorem 4. *A Markov chain is irreducible if and only if for all $j \in \{1, \ldots, n\}$ there exists a limit*

$$\lim_{t\to\infty} p_{ij}^{(t)} = p_j, \; i, j \in \{1, \ldots, n\},$$

independent of i, where p_j, $j \in \{1,\ldots,n\}$, form a unique solution of the following system of equations

$$\begin{cases} \sum_{i=1}^{n} p_i p_{ij} = p_j,\ j \in \{1,\ldots,n\} \\ \sum_{j=1}^{n} p_j = 1. \end{cases}$$

A special case of irreducible Markov chain is a regular Markov chain.

Definition 16. *A irreducible Markov chain is called regular if there exists $k \in \mathbb{N}$ such that all entries of the matrix Π^k are positive. In other words there exists $k \in \mathbb{N}$ such that from any state we can reach any state in exactly k steps.*

Definition 17. *Let Π be a transition matrix of a Markov chain. A stationary probability vector is a stochastic vector (see Definition 1) such that $\pi = \pi\Pi$. In other words π is a stochastic eigenvector associated with eigenvalue $\lambda = 1$ of matrix Π.*

Theorem 5 (see [15])**.** *Let $\Pi^{(m)} = (p_{ij}^{(m)})_{i,j\in\mathcal{S}}$ be a transition matrix in m steps of an irreducible aperiodic Markov chain (Z_n) with finite state space $\mathcal{S}$. Then*

(i) *for any $i, j \in \mathcal{S}$ there exists a limit $\lim_{m\to\infty} p_{ij}^{(m)} = \pi_j$, where $\pi_j > 0$,*
(ii) *Markov chain (Z_n) is recurrent,*
(iii) *a vector $\pi = (\pi_j)_{j\in\mathcal{S}}$ is a unique stationary probability vector, moreover, $\pi_j = \frac{1}{\mu_j}$ where μ_j is an average number of steps needed to reappear in state j.*

For regular Markov chains we have the following result.

Theorem 6. *Let Π be a transition matrix of an irreducible aperiodic Markov chain with finite state space. Then matrix $\Pi^{(m)}$ converges to a positive stochastic matrix W such that if π is a row of matrix W, then $\pi = \pi\Pi$.*

2.3. Analysis of Alpha 1,3-Galactosyltransferase 2

We show that a time homogeneous Markov chain is an appropriate simple model of Alpha 1,3-Galactosyltransferase 2 (A3GALT2). We use transition matrices as a criterion for identifying similarities in structure of this particular gene. A3GALT2 is a Protein Coding gene (a region of DNA) located in chromosome 1, position 33,306,766, consisting of 14,333 bases [16].

Let $\Omega = \{S_1, S_2, S_3, S_4\}$, where $S_1 = \text{A}$, $S_2 = \text{C}$, $S_3 = \text{G}$, $S_4 = \text{T}$, and $\mathcal{S} = \{1,2,3,4\}$ be a corresponding state space. We form a stochastic matrix (7) as follows. For example we would like to know how probable is that after adenine (A) occurs cytosine (C). We count all occurrences of a pair AC in gene A3GALT2 and divide it by number of all occurring pairs which start from A. Number of all such pairs is equal to number of occurrences of A provided that A is not the last nucleotid base in gene A3GALT2.

$$\Pi = \begin{array}{c} \\ \text{A} \\ \text{C} \\ \text{G} \\ \text{T} \end{array} \begin{array}{c} \begin{array}{cccc} \text{A} & \text{C} & \text{G} & \text{T} \end{array} \\ \begin{pmatrix} \frac{769}{3195} & \frac{745}{3195} & \frac{1093}{3195} & \frac{588}{3195} \\ \frac{1140}{4052} & \frac{1392}{4052} & \frac{350}{4052} & \frac{1170}{4052} \\ \frac{811}{3715} & \frac{934}{3715} & \frac{1254}{3715} & \frac{716}{3715} \\ \frac{475}{3370} & \frac{980}{3370} & \frac{1018}{3370} & \frac{897}{3370} \end{pmatrix} \end{array} \tag{7}$$

Note that because the last nucleotide base in gene A3GALT2 is T, in the denominator in last row we have 3370 instead of 3371.

Remark 12. *One can pose a question of biological interpretation of matrix Π. Does the occurrences of nucleotid bases can be used to identify a specific gene or, in general case, to identify an individual?*

Because all entries of matrix Π are positive, in the case of gene A3GALT2 a suitable Markov chain is irreducible (see Definition 10) and each $d_i = 1$ for $i \in \mathcal{S}$ thus our chain is aperiodic (see Definition 14). From Theorem 5 for any $i, j \in \mathcal{S}$ there exists a limit $\lim_{m\to\infty} p_{ij}^{(m)} = \pi_j > 0$, suitable Markov chain is recurrent, $\pi = (\pi_j)_{j\in\mathcal{S}}$ is a unique stationary probability vector and $\pi_j = \frac{1}{\mu_j}$ where μ_j is an average number of steps needed to reappear in state j.

Remark 13. *The stationary probability vector of matrix* Π *is*

$$\pi = (0.22292, 0.28265, 0.25923, 0.23521).$$

Remark 14. *Note that if nucleotide bases in gene A3GALT2 are a good estimation of a possible sequence of values of a Markov chain, then from Remark 13 probability of occurrence of a should be 0.22292, c: 0.28265, g: 0.25923 and t: 0.23521. Comparing this with computed probabilities from Table 1 (a: 0.222912, c: 0.282704, g: 0.259192, t: 0.235192) we see that they are correct up to the fourth decimal place.*

Table 1. Occurence of nucleotide bases in A3GALT2.

Nucleotide Bases	Occurence in A3GALT2	Probability (Occurence Divided by Length of Gene)
A	3195	0.222912
C	4052	0.282704
G	3715	0.259192
T	3371	0.235192

Corollary 1. *From Remark 13 we can compute approximate values of* μ_j*:* $\mu_1 \approx 4.486$, $\mu_2 \approx 3.538$, $\mu_3 \approx 3.858$, $\mu_4 \approx 4.252$ *which means that an average number of steps needed for each nucleotide base to reappear in our gene is approximately 4 for all bases. Thus we conclude that bases are uniformly distributed in gene A3GALT2 which means that they appear to be random and disorganized.*

All of the above considerations can be repeated for pairs of bases, see Table 2.

Table 2. Occurence of nucleotide pairs of bases in A3GALT2.

Pairs of Bases	Occurence in A3GALT2	Probability	Pairs of Bases	Occurence in A3GALT2	Probability
AA	769	$\frac{769}{14,332}$	GA	811	$\frac{811}{14,332}$
AC	745	$\frac{745}{14,332}$	GC	934	$\frac{934}{14,332}$
AG	1093	$\frac{1093}{14,332}$	GG	1254	$\frac{1254}{14,332}$
AT	588	$\frac{588}{14,332}$	GT	716	$\frac{716}{14,332}$
CA	1140	$\frac{1140}{14,332}$	TA	475	$\frac{475}{14,332}$
CC	1392	$\frac{1392}{14,332}$	TC	980	$\frac{980}{14,332}$
CG	350	$\frac{350}{14,332}$	TG	1018	$\frac{1018}{14,332}$
CT	1170	$\frac{1170}{14,332}$	TT	897	$\frac{897}{14,332}$

3. The Markov Process Model of Nucleotide Substitution

We assume that nucleotide substitution is follow a homogeneous Markov process. We take $\Omega = \{S_1, S_2, S_3, S_4\}$, where $S_1 = \mathrm{A}$, $S_2 = \mathrm{C}$, $S_3 = \mathrm{G}$, $S_4 = \mathrm{T}$. Let $\mathcal{S} = \{1, 2, 3, 4\}$ be a corresponding state space. Let $P(t) = \{P_{\mu\nu}(t)\}$ is a matrix of transition probabilities in time t. We assume that $P'(t) = QP(t)$ where $Q = \{Q_{\mu\nu}\}$ is the rate matrix of the process.

Remark 15. *Figures 2–5 show four situations in which the sequence starts with adenine, cytosine, guanine, and thymine, respectively. In each case, the probabilities of occurence of a given base stabilize. Finally, they converge to the probabilities forming a stationary probability vector. Note that, as predicted by Corollary 1, in each case stabilization is achieved after about four steps.*

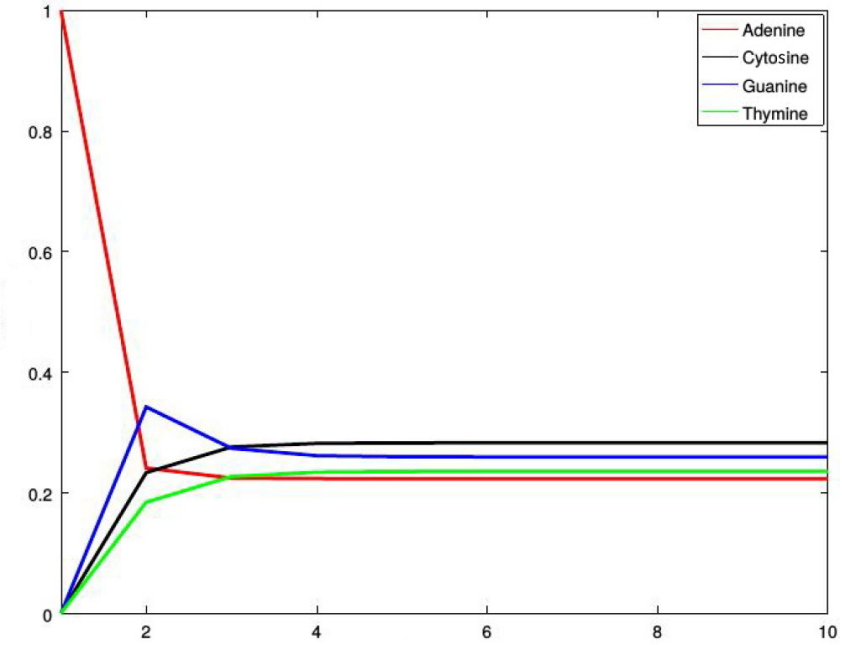

Figure 2. Probabilities of occurrences of the bases in consecutive steps starting from adenine.

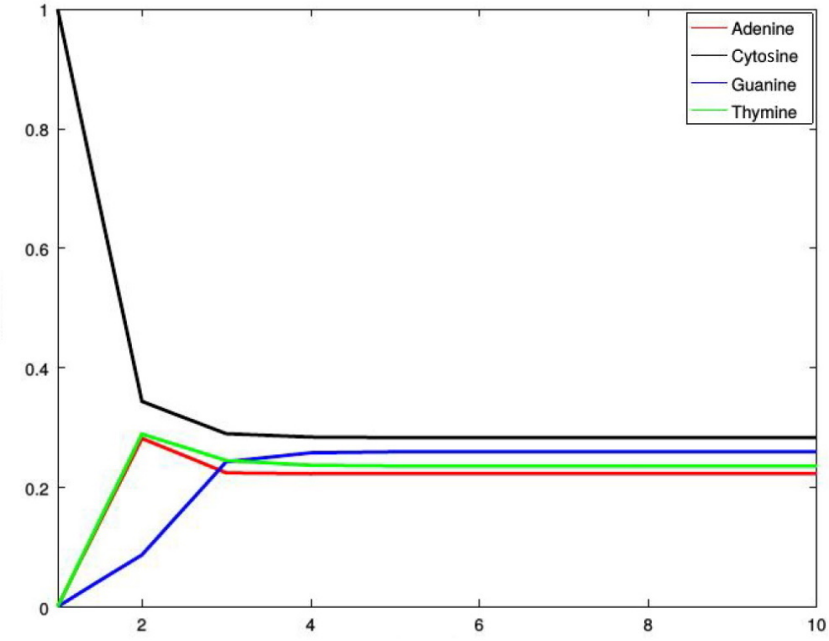

Figure 3. Probabilities of occurrences of the bases in consecutive steps starting from cytosine.

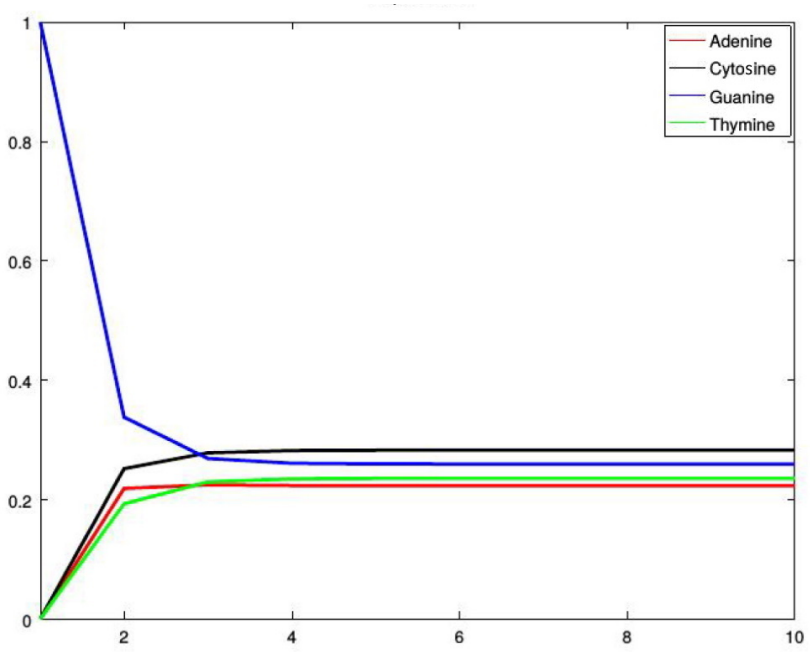

Figure 4. Probabilities of occurrences of the bases in consecutive steps starting from guanine.

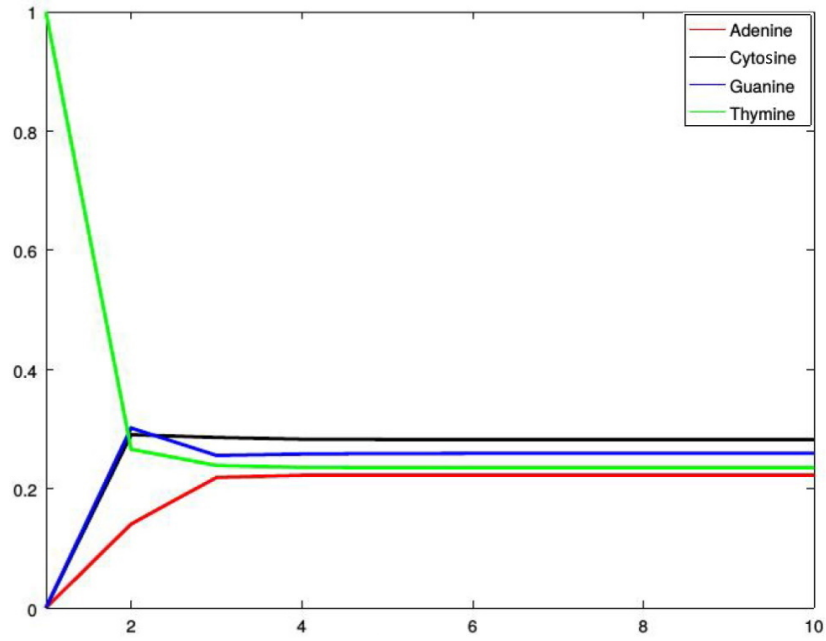

Figure 5. Probabilities of occurrences of the bases in consecutive steps starting from thymine.

4. Application in DNA Sequencing, Redesigned DNA

The meaning of DNA sequencing is here deemed to cover all methods used to determine the order of nucleotides along a DNA strand. The objective of this section is to present an example of applying Markov chains to complete short fragments of a DNA strand. Markov chains have been applied as mathematical models of real-life processes. Such real-life dynamical systems, examined with Markov-chain method, include

- queues of passengers arriving at an airport,
- currency exchange rates or
- animal population dynamics.

Markov chains are also applied to build algorithms calculating the PageRank value for a website (see [17]). The website PageRank value reflects the probability that a random internet surfer will land on this webpage upon clicking a link. Markov processes of various orders are used to model DNA sequencing (see also [6]). A Markov process of order m is one for which the probability of any event depends exclusively on the m preceding events. Statistically, DNA does not have the features of a first-order Markov chain. Higher-order models have been proposed for analyzing interrelations within a DNA sequence, see [6]. Statistical tests used in [6] properly determined the order of the Markov chain being tested for sequences of length 2^9 base pairs or higher. Nevertheless, the authors consider it important to present the method described below for local problems, that is for very short DNA sequences (54 base pairs in our example). We believe it is worth examining the results of local DNA completion based on the properties of first-order Markov chains, when the length of a DNA sequence is less than 2^9 base pairs. The method presented below is a simple, local tool for completing short DNA segments. The method has also educational value. Moreover, instead of analyzing a DNA sequence of the five unit nucleotides, we may use first-order Markov chains to analyze codons or, more precisely, amino acids encoded with those codons. It is worth recalling here some basic facts and conventions:

- Codon is a sequence of three nucleotides (a triplet) occurring in mRNA, a unit encoding a specific amino acid during protein synthesis;
- Proteins are built of 20 different amino acids;
- The sequence of amino acids in a protein exactly follows the sequence of the relevant codons in mRNA;
- Most amino acids are encoded in several ways (with different codons, which, however, differ from one another usually on the third place in the triplet only); owing to this, certain changes in the genetic information (mutations) do not affect the amino acid sequence;
- There are 61 codons encoding amino acids and 3 non-encoding codons (they are STOP codons: UAG, UAA, UGA); all in all: 4^3 various triplets;
- The AUG codon, read as the first one in mRNA by a ribosome during protein synthesis is known as the initiation or start codon;
- Since a mutation of a single nucleotide changes a single amino acid, the genetic code has to be read as non-overlapping, i.e., any given codon may be followed by any other codon;
- To get the form typical of DNA, each U in an mRNA codon should be replaced by T; for instance, TAA is the DNA equivalent of the mRNA codon UAA;
- In the case of a sequence of amino acids, understood as resulting from first-order Markov process action, the transition matrix is a square matrix of degree at most 21. One state is reserved for the three STOP codons, which do not encode amino acids.

Example 2. *Let us consider the human SATB1 gene, which, as research has revealed, is a major growth factor for breast cancer, see [18]. Let us generate a DNA sequence based on SATB1 (this gene is on chromosome 3, locus p23, on the minus strand). The table below sets forth a* 54*-base-pair-long segment of SATB1, position 18,389,139. Data is sourced from website [19], accessed upon entering human gene SATB1.*

The relevant state space comprises four nucleotides: $\Omega_1 = \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\}$. *The corresponding amino acid sequence is:*

Val Lys Arg Leu Ser Asp Lys Asn Lys Ser Ser Leu STOP Gln Leu Cys Cys STOP.

We give another sequence in Table 3, as a variation of the method consists in examining the sequence of amino acids and not the DNA sequence of base pairs. For the sequence of the DNA segment under this analysis, the state space is:

$$\Omega_2 = \{\text{Asn, Asp, Arg, Cys, Leu, Lys, Gln, Ser, Val, STOP}\}.$$

In the application of the method described below, it is important that in both tables the last element occurs at least twice. Assume that in the sequence in Table 4, the TCC codon (corresponding to the amino acid serine (Ser)) is missing. We want to properly complete the following sequence including three adjacent gaps:

$$\begin{array}{r} \text{GTCAAAAGACTCTCCGACAAAAACAAA}\square\square\square\text{AGTCTC} \\ \text{TAGCAGTTATGTTGTTAG} \end{array} \tag{8}$$

In the other approach, using representation with amino acids (see Table 3), we want to complete the following corresponding sequence including a single gap:

$$\text{Val Lys Arg Leu Ser Asp Lys Asn Lys } \square \text{ Ser Leu STOP Gln Leu Cys Cys STOP.} \tag{9}$$

The (extensive) transition matrix corresponding to sequence (8) is:

$$\Pi_1 = \begin{array}{c} \\ \text{A} \\ \text{C} \\ \text{G} \\ \text{T} \end{array} \begin{array}{c} \begin{array}{cccc} \text{A} & \text{C} & \text{G} & \text{T} \end{array} \\ \begin{pmatrix} \frac{9}{18} & \frac{3}{18} & \frac{5}{18} & \frac{1}{18} \\ \frac{4}{10} & \frac{1}{10} & \frac{1}{10} & \frac{4}{10} \\ \frac{2}{8} & \frac{1}{8} & \frac{0}{8} & \frac{5}{8} \\ \frac{3}{13} & \frac{5}{13} & \frac{2}{13} & \frac{3}{13} \end{pmatrix} \end{array} \tag{10}$$

Let us observe that the number of occurrences of G *in sequence (8) is* 8*, as we do not count the last occurrence, because it is not paired. The number of occurrences of* A *in sequence (8) is* 18*, as we do not count the last occurrence of* A*, before the lacking fragment, because this occurrence is not paired. Analogously, the transition matrix corresponding to sequence (9) is:*

$$\Pi_2 = \begin{array}{c} \\ \text{Asn} \\ \text{Asp} \\ \text{Arg} \\ \text{Cys} \\ \text{Leu} \\ \text{Lys} \\ \text{Gln} \\ \text{Ser} \\ \text{Val} \\ \text{Stop} \end{array} \begin{array}{c} \begin{array}{cccccccccc} \text{Asn} & \text{Asp} & \text{Arg} & \text{Cys} & \text{Leu} & \text{Lys} & \text{Gln} & \text{Ser} & \text{Val} & \text{Stop} \end{array} \\ \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ \frac{1}{3} & 0 & \frac{1}{3} & 0 & 0 & 0 & 0 & \frac{1}{3} & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{2} & 0 & 0 & 0 & 0 & 0 & \frac{1}{2} \\ 0 & 0 & 0 & \frac{1}{3} & 0 & 0 & 0 & \frac{1}{3} & 0 & \frac{1}{3} \\ \frac{1}{3} & 0 & \frac{1}{3} & 0 & 0 & 0 & 0 & \frac{1}{3} & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{3} & 0 & 0 & \frac{1}{3} & 0 & 0 & \frac{1}{3} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \end{pmatrix} \end{array} \tag{11}$$

Note that, after rounding, the stationary probability vector of matrix Π_1 *is equal to:*

$$\pi_1 \approx (0.367347, 0.204082, 0.163265, 0.265306).$$

For the sake of comparison, we below give rounded relative frequencies of nucleotide occurrences, as disclosed in Table 4:

$$\text{A} : 0.351852, \text{C} : 0.222222, \text{G} : 0.166667, \text{T} : 0.259259.$$

The values sourced from Table 4 are not identical to the respective coordinates of the vector π_1*, given that:*

1. *the sequence of* 54 *nucleotides is short,*
2. *in sequence (8), three nucleotides are lacking (cf. Remark 14).*

Using the transition matrix Π_1*, we will run an experiment, described below, which will allow us to complete sequence (8). One can proceed analogously using the matrix* Π_2*, which we will not do, given the symmetry of the method. The description of the experiment makes it possible to repeat it with no IT tools.*

Prepare four boxes, labeled A, C, G *and* T. *In each box, there are assorted balls labeled* A, C, G *and* T. *The numbers of balls of individual colors in box* A *are in proportion to respective entries of the first row of the matrix* Π_1*. Thus there are* 9 *balls labeled* A, 3 *balls labeled* C, 5 *balls labeled* G, *and* 1 *ball labeled* T, *a total of* 18 *balls. We fill the other boxes* (C, G *and* T) *analogously. Now the experiment begins. In sequence (8), there is a gap after* A. *Therefore, we draw one ball from box* A *at random. Assume we have drawn* T, *which can be done with probability* $\frac{1}{18}$*. In the next step we draw from the box labeled the same way as the most recently drawn ball; in our experiment it is box* T. *Assume we have drawn* C *from box* T, *which can be done with probability* $\frac{5}{13}$*. Proceeding this way, we now draw a ball from box* C. *Assume we have again drawn* C, *which can be done with probability* $\frac{1}{10}$*. Thus we have generated three consecutive elements of the sequence* GTCAAAAGACTCTCCGACAAAAACAAA, *namely* TCC, *and we fill the gap in sequence (8) with this result. We have recovered the original sequence presented in Table 4. The algorithm works as shown in Figure 6.*

For the reader's convenience, we present in Figure 7 below a computer program written in the Python 3 language, the source code is also available in GitHub [20].

Table 3. The sequence of amino acids corresponding to 54-base-pair-long segment of SATB1.

1,2,3	4,5,6	7,8,9	10,11,12	13,14,15	16,17,18	19,20,21	22,23,24	25,26,27
Val	Lys	Arg	Leu	Ser	Asp	Lys	Asn	Lys
28,29,30	31,32,33	34,35,36	37,38,39	40,41,42	43,44,45	46,47,48	49,50,51	52,53,54
Ser	Ser	Leu	STOP	Gln	Leu	Cys	Cys	STOP

Table 4. Numbered 54-base-pair-long segment of SATB1.

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
G	T	C	A	A	A	A	G	A	C	T	C	T	C	C	G	A	C
19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36
A	A	A	A	A	C	A	A	A	T	C	C	A	G	T	C	T	C
37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54
T	A	G	C	A	G	T	T	A	T	G	T	T	G	T	T	A	G

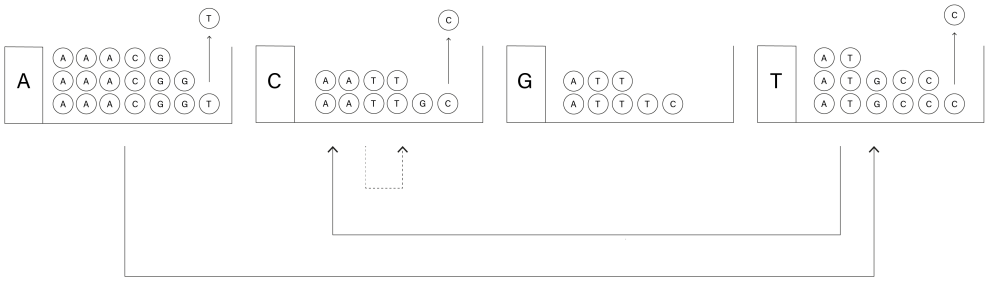

Figure 6. Illustration for Example 2.

```
from scipy import stats

text = ['A','C','G','T']

data = (1,2,3,4)
pA = (9/18,3/18,5/18,1/18)
pC = (4/10,1/10,1/10,4/10)
pG = (2/8,1/8,0/8,5/8)
pT = (3/13,5/13,2/13,3/13)

A = stats.rv_discrete(values=(data,pA))
C = stats.rv_discrete(values=(data,pC))
G = stats.rv_discrete(values=(data,pG))
T = stats.rv_discrete(values=(data,pT))

def gen(start):
    res = []
    last = start
    for i in range(3):
        if last==1: last=A.rvs()
        elif last==2: last=C.rvs()
        elif last==3: last=G.rvs()
        else: last=T.rvs()
        res.append(last)
    return res

last_letter = 1

print('{}{}{}'.format(*[text[i-1] for i in gen(last_letter)]))
```

Figure 7. Program executes computations from Example 2 with probabilities from the transition matrix Π_1 and generates the completion of sequence (8) with exactly three elements.

5. Conclusions

The method just presented is primarily of educational value. The procedure is described suggestively and, the authors believe, explanatory, which makes it possible for the method to be used in a more general context, also by non-mathematicians. The authors do not imply that the probability of achieving the proper completion of a DNA genome is satisfactory; they only present a tool which may be used for such completion and with which they would like to familiarize the reader. The authors are aware that the contemporary efforts in the area of DNA genome completion are focusing on deep learning rather than on Markov chains even of higher orders, see [8]. This paper proposes an alternative tool that can be explained suggestively and deeply. It is now clear that the deep learning methods lead to more exact completions than the Markov chains methods. Yet, our method allows for understanding of what happens behind the process of proper completion and sequencing. It should therefore be treated as of explanatory and educational value, with a potential for future research. It is worth asking, if the algorithms presented in Example 2 might be used to easily generate test data for more advanced deep learning algorithms (see [21]).

Author Contributions: Conceptualization, M.Z. (Maciej Zakarczemny); methodology, M.Z. (Maciej Zakarczemny) and M.Z. (Małgorzata Zajęcka); software, M.Z. (Małgorzata Zajęcka); validation, M.Z. (Maciej Zakarczemny) and M.Z. (Małgorzata Zajęcka); formal analysis, M.Z. (Maciej Zakarczemny) and M.Z. (Małgorzata Zajęcka); investigation, M.Z. (Maciej Zakarczemny) and M.Z. (Małgorzata Zajęcka); resources, M.Z. (Maciej Zakarczemny) and M.Z. (Małgorzata Zajęcka); data curation, M.Z. (Maciej Zakarczemny) and M.Z. (Małgorzata Zajęcka); writing—original draft preparation, M.Z. (Maciej Zakarczemny) and M.Z. (Małgorzata Zajęcka); writing—review and editing, M.Z. (Maciej Zakarczemny) and M.Z. (Małgorzata Zajęcka); visualization, M.Z. (Maciej Zakarczemny) and M.Z. (Małgorzata Zajęcka); supervision, M.Z. (Maciej Zakarczemny) and M.Z. (Małgorzata Zajęcka). All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Blaisdell, B.E. Markov chain analysis finds a significant influence of neighboring bases on the occurrence of a base in eucaryotic nuclear DNA sequences both protein-coding and noncoding. *J. Mol. Evol.* **1985**, *21*, 278–288. [CrossRef] [PubMed]
2. Brendel, V.; Beckmann, J.S.; Trifonov, E.N. Linguistics of nucleotide sequences: Morphology and comparison of vocabularies. *J. Biomol. Struct. Dyn.* **1986**, *4*, 11–21. [CrossRef] [PubMed]
3. Avery P.J.; Henderson D.A. Fitting Markov chain models to discrete state series such as DNA sequences. *Appl. Stat.* **1999**, *48*, 53–61. [CrossRef]
4. Wu, T.-J.; Hsieh, Y.-C.; Li, L.-A. Statistical Measures of DNA Sequence Dissimilarity under Markov Chain Models of Base Composition. *Biometrics* **2001**, *57*, 441–448. [CrossRef] [PubMed]
5. Pérez-Lechuga, G.; Venegas-Martínez, F.; Martínez-Sánchez, J.F. Mathematical Modeling of Manufacturing Lines with Distribution by Process: A Markov Chain Approach. *Mathematics* **2021**, *9*, 3269. [CrossRef]
6. Usotskaya, N.; Ryabko, B. Applications of information-theoretic tests for analysis of DNA sequences based on Markov chain models. *Comput. Stat. Data Anal.* **2009**, *53*, 1861–1872. [CrossRef]
7. Ching, W.; Huang, X.; Ng, M.K.; Siu, T.K. *Markov Chains: Models, Algorithms and Applications*; Springer: New York, NY, USA, 2013; Volume 189.
8. Yang, A.; Zhang, W.; Wang, J.; Yang, K.; Zhang, L. Review on the Application of Machine Learning Algorithms in the Sequence Data Mining of DNA. *Front. Bioeng. Biotechnol.* **2020**, *8*, 1032. [CrossRef] [PubMed]
9. Bell, C. Algorithmic music composition using dynamic markov chains and genetic algorithms. *J. Comput. Sci. Coll.* **2011**, *27*, 99–107.
10. Linskens, E.J. Music Improvisation Using Markov Chains. Available online: https://dke.maastrichtuniversity.nl/gm.schoenmakers/wp-content/uploads/2015/09/Linskens-Final-Draft.pdf (accessed on 16 January 2022).
11. Liu, Y.W.; Selfridge-Field, E. Modeling Music as Markov Chains: Composer Identification. 2002. Available online: https://ccrma.stanford.edu/~jacobliu/254report/ (accessed on 16 January 2022).
12. Hayes, B. First Links in the Markov Chain. *Am. Sci.* **2013**, *101*, 92. Available online: https://www.americanscientist.org/sites/americanscientist.org/files/201321152149545-2013-03Hayes.pdf (accessed on 16 January 2022). [CrossRef]
13. Markov, A. An Example of Statistical Investigation of the Text Eugene Onegin Concerning the Connection of Samples in Chains. *Sci. Context* **2006**, *19*, 591–600. [CrossRef]
14. Yang, Z.; Jin, S.; Huang, Y.; Zhang, Y.; Li, H. Automatically generate steganographic text based on markov model and huffman coding. *arXiv* **2018**, arXiv:1811.04720.
15. Jakubowski, J.; Sztencel, R. *Introduction to Probability Theory*; Script: Poland, Warsaw, 2010. (In Polish)
16. A3GALT2 at National Center for Biotechnology Information, U.S. National Library of Medicine. Available online: https://www.ncbi.nlm.nih.gov/gene/127550 (accessed on 16 January 2022).
17. PageRank, in Wikipedia. Available online: https://en.wikipedia.org/wiki/PageRank (accessed on 16 January 2022).
18. Wang, X.; Yu, X.; Wang, Q.; Lu, Y.; Chen, H. Expression and clinical significance of SATB1 and TLR4 in breast cancer. *Oncol. Lett.* **2017**, *14*, 3611–3636. [CrossRef] [PubMed]
19. Wolfram Alpha LLC. Wolfram | Alpha. Available online: https://www.wolframalpha.com/ (accessed on 16 January 2022).
20. Zakarczemny, M. Note on DNA Analysis and Redesigning Using Markov Chain Simulations. Available online: https://github.com/MaciejZakar/SATB1program (accessed on 16 January 2022).
21. Jääskinen, V. *Bayesian Stochastic Partition Models For Markovian Dependence Structures*; University of Helsinki: Helsinki, Finland, 2015.