

Применение структур случайных последовательностей для описания свойств мтДНК и определения принадлежности отдельных мтДНК к их хозяйской группе животных.

Филатов О.В.

Филатов Олег Владимирович - инженер-программист,
ЗАО «Научно технический центр «Модуль», г. Москва, fil_post@rambler.ru

Аннотация: исследование структуры мтДНК с позиций Комбинаторики Длинных Последовательностей (КДП) показало, что числовые характеристики логических образований (составные события, цуги) в мтДНК в значительной степени равны аналогичным характеристикам в случайной пос-ти. Однако, исследование мтДНК методами КДП выявило необходимость ввести новый вид цуг в КДП, который получил название «Однобуквенные цуги» - $C(1L)$. В этой статье даются основные формулы для $C(1L)$ цуг и показано их применение при исследовании структуры мтДНК, а так же показано, как при помощи описанных в КДП структур, которые так же присущи и мтДНК благодаря их родству со случайными последовательностями, можно классифицировать (распределять) мтДНК по группам животных: Amphibians; Birds; Fishes; Mammals; Reptiles; Fungi; Insects; Land Plants; Roundworms.

Ключевые слова: мтДНК, составные события, цуги, КДП, комбинаторика.

The use of random sequence structures to describe the properties of mtDNA and to determine whether individual mtDNAs belong to their host group of animals.

Filatov O.V.

Filatov Oleg Vladimirovich - Software Engineer,
SCIENTIFIC AND TECHNICAL CENTER «МОДУЛЬ», MOSCOW

Abstract: *a study of the structure of mtDNA from the position of the Combinatorics of Long Sequences (CDP) showed that the numerical characteristics of logical formations (compound events, trains) in mtDNA are substantially equal to similar characteristics in a random number. However, the study of mtDNA by CDP methods revealed the need to introduce a new type of train in CDP, which was called "Single-letter trains" - C (1L). This article gives the basic formulas for the C (1L) train and shows their application in the study of the structure of mtDNA, and also shows how, using the structures described in CDP, which are also inherent in mtDNA due to their relationship to random sequences, it is possible to classify (to distribute) mtDNA according to animal groups: Amphibians; Birds Fishes; Mammals; Reptiles Fungi; Insects Land plants; Roundworms.*

Keywords: *mtDNA, compound events, trains, CDPs, combinatorics.*

УДК: 51; 34.23.37

Введение

Сейчас ДНК представляют в виде хаотичноподобной последовательности из четырёх букв. Исследования установили, что отклонения характеристик ДНК последовательности от соответствующих характеристик случайной пос-ти не являются значительными. Поэтому логично описывать, по крайней мере, мтДНК пос-ти в терминах, которые характеризуют случайные математические пос-ти. Основные характеристики случайных пос-ей были выявлены при их компьютерном исследовании и моделировании. Все полученные характеристики структуры случайных пос-ей были оформлены в виде комбинаторного направления – «Комбинаторика Длинных Последовательностей» (КДП). На основе структурных элементов, которые КДП выделяет в случайных последовательностях, были сформированы формулы, которые в данной работе называются фильтрами, так как они выполняют работу по выделению нужных структурных

особенностей мтДНК и сокрытию не нужных в данный момент структур мтДНК. Применение фильтров позволяет не только обнаружить общие характеристики мтДНК в отдельных группах животных, но и определять принадлежность мтДНК неизвестного животного к одной из известных групп живых существ. Пример такой идентификации секвенированного мтДНК неизвестного животного, к одной из существующих групп животных, приведен в этой статье.

В статье используется информационный пул более чем из полутысячи секвенированных мтДНК. В результате КДП обработки всех этих мтДНК была сформирована База Данных (БД) значительно меньшего размера, чем исходная БД, в которой мтДНК представлены в виде последовательностей букв. Созданная КДП БД заменила описание всех мтДНК буквенными последовательностями на числовые КДП характеристики мтДНК. Эти числовые КДП характеристики мтДНК были получены в результате работы компьютерной программы рассчитавшей КДП величины для каждого живого существа и граничные, предельные величины, для каждой группы животных [7]. То есть, каждый числовой диапазон, характеризующий особенности принадлежащих к нему группы живых существ, образован по результатам обработки 40 - 43 мтДНК живых существ, принадлежащих одной из групп: *Amphibians; Birds; Fishes; Mammals; Reptiles; Fungi; Insects; Land Plants; Roundworms*.

Основная часть

Не разрывные цуги (№ 1) из составных событий равной длины.

В случайных равновероятных пос-тях (число равновероятных исходов последовательности обозначим буквой V) известны логические образования – цуги [1, 2, 3, 5, 6]. Цуги образуют из составных событий равной длины [1, 5, 6] (длина составного события обозначим буквой n). Цуги формируются из составных событий равной длины, когда составные события следуют друг за

другом (число составных событий входящих в цугу обозначим буквой w). В цугах первого вида (№1), следующие друг за другом составные события, образованы из разных элементарных событий [1]. Примеры цуг из составных событий равной длины, образованных из четырёх букв (то, что цуга может быть образована из четырёх букв, обозначается как «4L»): $A; C; G; T$:
 $\overset{n=2}{\underset{v=4}{C}}(4L)_{w=5} = \text{«GGAACCAATT»}$. В этом примере: $\overset{n=1}{\underset{v=4}{C}}(4L)_{w=10} = \text{«ACTCTCTACA»}$ - цуга образована не из полного набора букв (не из четырёх букв, а из трёх букв), в цуге нет буквы «G» (буква «G» присутствует в случайной пос-ти, но в эту конкретную цугу не входит).

Разрывные однобуквенные цуги (№ 2) из составных событий равной длины.

В качестве логической основы формирования цуг можно положить и другой принцип – принцип разрывного выпадения составных событий равной длины, причём составные события образующие цугу второго вида (№2) состоят из одинаковых элементарных событий, в отличие от цуг вида №1.

Под разрывностью цуг №2 понимается следующее: составные события, объединяемые логическим образом в цугу образуются из элементарных событий одного типа; после окончания предшествующего цугового составного события и перед началом последующего цугового составного события располагаются элементарные события не равные элементарным событиям образующим цугу. Причем, эти элементарные события располагаются между составными событиями цуги в любых количествах, но их численность всегда больше нуля. Комбинации и сочетания элементарных событий, лежащих между составными событиями цуги, могут быть любые.

То, что цуга образована одним символом, сокращенно указывается через число «1» и букву «L»: «1L», где: «1» - это указание числа букв

образующих цугу, а «L» - первая буква английского слова Letter, то есть «1L» - это «одна буква».

Примеры образования из элементарных событий: $A; C; G; T$ и составных событий равной длины, однобуквенных цуг $C(1L)$, составные события в цугах выделены жирным шрифтом и подчёркнуты:

Цуга $C(4L)$ из буквы «A»: $\frac{A(1)}{V=4}C(1L)_{w=4} = \langle \dots AACACCCGTAGATTCAA \dots \rangle;$

Цуга $C(4L)$ из буквы «C»: $\frac{C(1)}{V=4}C(1L)_{w=1} = \langle \dots ACACCCTAGATTCCA \dots \rangle;$

Цуга $C(4L)$ из буквы «G»: $\frac{G(1)}{V=4}C(1L)_{w=3} = \langle \dots AGAGGTAGGAGGAAGGGT \dots \rangle;$

Цуга $C(4L)$ из буквы «T»: $\frac{T(1)}{V=4}C(1L)_{w=2} = \langle \dots GTGTTTTTCTTTTTGTTG \dots \rangle.$

Далее речь пойдёт о «Нулевых цугах» $C0$. Понятие нулевых цуг $C0$ дано в работе [1, 5, 6].

Экспериментально установлено - суммарная численность всех четырёх нулевых цуг $C0(1L)$ случайной пос-ти из букв $A; C; G; T$ (элементарных событий), равна 30% от числа букв (элементарных исходов N), ф.1.1:

$$\frac{ACGT}{V=4}C0(1L)_N = C0(A) + C0(C) + C0(G) + C0(T) = 0,3 \cdot N \quad \text{Ф. 1.1}$$

В случайной последовательности из четырёх равновероятных исходов $V = 4$, математическое ожидание численности для любой цуги $\frac{X}{V=4}C0(1L)_N$, где $X \in \{A; C; G; T\}$, равно одной четверти от общего числа цуг пос-ти ($C0(A) = C0(C) = C0(G) = C0(T)$), ф.1.2:

$$\frac{X}{V=4}C0(1L)_N = \frac{\frac{ACGT}{V=4}C0_N}{4} = \frac{3 \cdot N}{40} \quad \text{Ф. 1.2}$$

Где N - число членов случайной равновероятной V последовательности.

Или : $\frac{ACGT}{V=4}C0(1L)_N = 4 \cdot \frac{X}{V}C0(1L)_{X;N} = 0,3 \cdot N$, где $X \in \{A; C; G; T\}$.

Найденная экспериментальным путём ф.1.3 позволяет рассчитать число цуг ${}_{V=4}^{X(n)}C0(1L)_N$, с базовой длиной n , как функцию от числа равновероятных исходов V и числа случайных событий N : ${}_{V=4}^{X(n)}C0(1L)_N = f(n, V, N)$:

$${}_{V=4}^{X(n)}C0(1L)_N = \frac{N}{V^n} \cdot \left(\frac{V-1}{V}\right)^2 \cdot \left(1 - \frac{1}{4^{n-1}} + \frac{1}{4^n}\right) \quad \text{Ф. 1.3}$$

В [1, 4, 5, 6] дана формула, которая описывает число составных событий ${}_V^nSX_{X;N}$ случайной пос-ти из V исходов (вершин), как функцию трёх параметров ${}_V^nSX_N = f(n, V, N) = \frac{N}{V^n} \cdot \left(\frac{V-1}{V}\right)^2$, учитывая это выразим ф.1.3 в виде ф.1.4 для разрывных однобуквенных цуг из составных событий:

$${}_{V=4}^{X(n)}C0(1L)_N = {}_V^nSX_N \cdot \left(1 - \frac{1}{4^{n-1}} + \frac{1}{4^n}\right) \quad \text{Ф. 1.4}$$

Подставляя в ф.1.3 вместо V его значение ($V=4$) получим выражение: $\frac{9 \cdot N}{16} \cdot \left(\frac{1}{4^n} - \frac{1}{4^{2n-1}} + \frac{1}{4^{2n}}\right)$. Замечаем, что: $\frac{1}{4^n} = \left(\frac{1}{4}\right)^n = p^n$, где p - вероятность реализации одной из четырёх равновероятной возможности. Заменим в ф.1.3 все $\left(\frac{1}{4}\right)^n$ на p^n , где $p = \frac{1}{4}$, получим ф.1.5:

$${}_{V=4}^{X(n)}C0(1L)_N = 9p^{n+2}(1 - 3p^n)N \quad \text{Ф. 1.5}$$

Где: $X \in \{A; C; G; T\}$ - «X» это одна из букв множества; N - число членов образующих случайную последовательность.

Вводя в ф.1.5 составные события [1,4,5,6]: ${}_V^nSX_N$, получим ф.1.6

$${}_{V=4}^{X(n)}C0(1L)_N = {}_V^nSX_N \cdot (1 + p^n - p^{n-1}) \cdot N \quad \text{Ф. 1.6}$$

Где: $p = \frac{1}{V} = \frac{1}{4}$.

Отметим, что суммы: не разрывных нулевых цуг ${}_VC0(4L)_N$, и суммы разрывных однобуквенных нулевых цуг ${}_{V=4}^{ACGT}C0(1L)_N$, равны, ф.1.7:

$${}_V C0(4L)_N = \frac{ACGT}{V=4} C0(1L)_N = 0,3 \cdot N \quad \text{Ф. 1.7}$$

Действительно: ${}_V C0(4L)_N = 4 \cdot \frac{X(n)}{V=4} C0(1L)_N = \frac{2 \cdot N}{V+1} \cdot \left(\frac{V-1}{V} \right) = \frac{2 \cdot N}{5} \cdot \frac{3}{4} = 0,3 \cdot N.$

Пояснение таблицы 1.

В таблице 1 представлен процесс определения принадлежности некоего имеющегося мтДНК к одной из имеющихся в базе данных групп животных.

В распоряжении исследователей есть файл с секвенированным мтДНК, в котором $N_{Exp}=18133$ буквы. В строке «B2» лежат значения мтДНК для которого надо найти группу животных к которой он принадлежит. Применяя к этому мтДНК, поочерёдно, разные фильтрующие формулы (фильтры: F1; F2; F3), мы выявляем те или иные свойства этого генома, что проявляется в виде изменения характеризующих его чисел, строка «B2», столбцы: 2, 3, 4.

В столбцах 2, 3, 4, таблицы 1, строки: «B3» - «B12» указаны максимальные и минимальные, граничные величины, которые присущи для десяти групп животных.

Нужно выявить попадание величин определяемого по группе мтДНК, строка «B2», в диапазоны соответствующих величин десяти групп животных, строки: «B3» - «B12». По результатам сравнений значений этих величин: определяемого по группе мтДНК, и, соответствующих величин принадлежащих группам животных, определим к какой группе животных это мтДНК («B2») может относиться. Если значение определяемого мтДНК («B2») не попадает внутрь диапазона [max – min] какой либо группы, то, мтДНК не принадлежит этой группе. В столбце 5 («Итоговый результат отбора»), в строке «B12», сделан вывод: «Insects - владельцы мтДНК» (вывод соответствует действительности, тестируемый мтДНК принадлежит насекомому *Sinochlora longifissa*).

Таблица 1. «Выявление группы животных, к которой принадлежит мтДНК».

		по результатам фильтрации меняются значения диапазонов групп животных			Итоговый результат отбора
	1	2	3	4	5
B1	Названия групп животных	F1- фильтр 1	F2-фильтр 2	F3-фильтр3	Владельцы мтДНК
B2	<u>Определяемый мтДНК</u>	<u>0,417391</u>	<u>1,566941</u>	<u>0,565400</u>	
B3	Roundworms. Значения в группе max, min.	0,231488 -0,94617	Не владелец	Не владелец	Не владелец
B4	Birds. Значения в группе max, min.	1,059845 0,743515	Не владелец	Не владелец	Не владелец
B5	Land Plants. Значения в группе max, min.	0,072352 -0,09702	Не владелец	Не владелец	Не владелец
B6	Fungi_Basidiomycetes. Значения в группе max, min.	0,019371 -0,47449	Не владелец	Не владелец	Не владелец
B7	Fungi_Ascomycetes. Значения в группе max, min.	0,167541 -0,26113	Не владелец	Не владелец	Не владелец
B8	Reptiles. Значения в группе max, min.	0,942704 0,504661	Не владелец	Не владелец	Не владелец
B9	Amphibians. Значения в группе max, min.	<u>0,839037</u> <u>0,262124</u>	1,534853 1,096302	Не владелец	Не владелец
B10	Mammals. Значения в группе max, min.	0,907136 0,428963	1,434449 0,858845	Не владелец	Не владелец
B11	Fishes. Значения в группе max, min.	0,871397 0,4284	<u>1,569587</u> <u>0,874985</u>	0,504765 0,408646	Не владелец
B12	Insects. Значения в группе max, min.	<u>0,559021</u> <u>-0,82703</u>	<u>2,209385</u> <u>0,807674</u>	<u>0,615471</u> <u>0,538235</u>	<u>Insects - владельцы мтДНК</u>
Btн221: file: C0(C) del tC0(X) - C0(G) del tC0(X).xlsx: Ф1: «C0(C) /tC0(X) - C0(G) / tC0(X)»; Ф2: C0(T)/tC0(X); Ф.3: C0(A)/SS(A) file № 43; Insects Sinochlora longifissa.txt_out.dat; $N_{Exp} = 18133$					

Результаты обработки мтДНК по фильтру 1.

Первый фильтр работает с экспериментально обнаруженными и, так же, теоретическими рассчитанными однобуквенными цугами (№2) образованными из букв «С» и «G» файла буквенной записи мтДНК. В таблице 1, в столбце 2 видно, что определяемый мтДНК не попал своим значением 0,417391 (полученное по результатам обработки мтДНК фильтром 1) внутрь диапазонов шести групп: Roundworms, Birds, Land Plants, Fungi_Basidiomycetes, Fungi_Ascomycetes, Reptiles – эти группы исключаем из дальнейшего рассматривания в качестве возможного включения в них искомого мтДНК.

Введём сокращённое обозначение применяемых фильтров через букву F и цифру за ней (порядковый номер фильтра). Тогда логическая запись первого фильтра, по которому было отброшено шесть групп животных из девяти, будет выглядеть так: $F1 = \langle C0(C) / tC0(X) - C0(G) / tC0(X) \rangle$ Запишем это подробнее: $F1 = \frac{\sum_{v=4}^C C0(1L)_{Exp}}{\sum_{v=4}^C C0(1L)_{Teor}} - \frac{\sum_{v=4}^G C0(1L)_{Exp}}{\sum_{v=4}^G C0(1L)_{Teor}} = 0,417391$.

Учитывая, что теоретическая величина $\sum_{v=4}^C C0(1L)_{Teor}$ по ф.1.2 равна $\frac{3 \cdot N}{40}$, получим формулу для $F1$ («Фильтра 1»), ф.2.1:

$$F1 = \frac{40}{3 \cdot N_{Exp}} \cdot \left(\sum_{n(C)=1}^{n(C)=max} \frac{C(n)}{V=4} C0(1L)_{Exp} - \sum_{n(G)=1}^{n(G)=max} \frac{G(n)}{V=4} C0(1L)_{Exp} \right) \quad \Phi. 2.1$$

Результаты обработки мтДНК по фильтру 2.

Второй фильтр работает с экспериментально обнаруженными и, так же, теоретическими рассчитанными однобуквенными цугами (№2) образованными из буквы «Т» файла буквенной записи мтДНК. В таблице 1, в столбце 3 видно, что определяемый мтДНК не попал своим значением 1,566941 (полученное по результатам обработки фильтра 2: $C0(T)/tC0(X)$) внутрь диапазонов двух групп: Amphibians и Mammals – эти группы исключим из дальнейшего рассмотрения в качестве возможного включения в них искомого мтДНК. В ф.2.2.1 приведена формула для расчёта значений по второму фильтру $F2$:

$$F2 = \sum_{n(T)=1}^{n(T)=max} \frac{T(n)}{V=4} C0(1L)_{Exp} : \sum_{n(X)=1}^{n(X) \gg 1} \frac{X(n)}{V=4} C0(1L)_{Teor} \quad \Phi. 2.2.1$$

Где: $X(n) = A(n) = A(n) = G(n) = T(n)$; индекс $Teor$ - обозначает, что значение каждой нулевой цуги рассчитывается теоретическим путём, то есть по ф.1.3 (а не берётся из экспериментальных данных, как в случае с Exp).

Учитывая, что по ф.1.2: $\sum_{n(X)=1}^{n(X) \gg 1} \sum_{V=4}^{X(n)} C0(1L)_{Teor} = \sum_{V=4}^X C0(1L)_N = \frac{3 \cdot N}{40}$,

перепишем ф.2.2.1 в виде ф.2.2.2:

$$F2 = \frac{40}{3 \cdot N_{Exp}} \cdot \sum_{n(T)=1}^{n(T)=max} \sum_{V=4}^{T(n)} C0(1L)_{Exp} \quad \text{Ф. 2.2.2}$$

Результаты обработки мтДНК по фильтру 3.

Третий фильтр «C0(A)эксп / SS(A)эксп» использует для своей работы при анализе исследуемого мтДНК экспериментально полученные в файле, фактические, численности цуг «C0(A)эксп» и составных событий «SS(A)эксп» [1, 4, 5, 6]. Впрочем, это касается не только исследуемой мтДНК, но и мтДНК всех животных во всех группах (Fishes, Insects), с которыми производится сравнение значения 0,565400 в столбце 4. В результате этого сравнения сравниваемая величина оказалась внутри граничных значений группы Insects: $0,615471 > \underline{0,565400} > 0,538235$. Это значит, что искомый мтДНК принадлежит одному из видов входящих в группу Insects. Формальная формула фильтра 3 приведена в ф.2.3:

$$F3 = \sum_{n(A)=1}^{n(A)=max} \sum_{V=4}^{A(n)} C0(1L)_{Exp} : \sum_{n(A)=1}^{n(A)=max} \sum_{V=4}^{A(n)} SS_{Exp} \quad \text{Ф. 2.3}$$

Обсуждение

Без применения законов Комбинаторики длинных пос-ей (КДП), мтДНК выглядят абсолютно хаотическим нагромождением нуклеотидов. Но при сравнении структур мтДНК со структурой случайной пос-ти, оказывается, что структуры мтДНК отличаются не более чем на 10% от хорошо описываемых формулами КДП структур случайных пос-ей. То есть у мтДНК высокая степень родства со случайными последовательностями, но, тем не менее, не смотря на высокий уровень сродства к хаосу, мтДНК несут в себе информацию о строении митохондрий и даже строении организмов, в которых они живут. То есть, у структур мтДНК есть особенность, которой

нет у структур случайных пос-ей. Структуры мтДНК группируются по признаку принадлежности своим группам животных. В статье было показано, как используя эту группировку можно классифицировать принадлежность не опознанного (не известного) мтДНК к определённой группе животных. Что, на самом деле, очень удивительно. Во первых, мтДНК не обязано нести в себе признаки своих хозяев. Во вторых, было бы нормально, если КДП характеристики мтДНК разных групп животных случайным образом перемешивались, а не группировались в разделяемые кластеры в зависимости от их принадлежности к той или иной форме жизни их хозяев.

Выводы

- 1) В КДП характеристиках мтДНК последовательностей существуют отклонения от аналогичных КДП характеристик случайных пос-тей.
- 2) Благодаря существованию уникального (неповторимого) набора КДП характеристик, являющихся следствием отклонений от характеристик случайных пос-тей, для каждого мтДНК, каждое мтДНК можно однозначно идентифицировать набором КДП характеристик.
- 3) Размеры любого набора КДП характеристик для любого мтДНК, выраженного в байтах, в десятки раз меньше размера буквенной записи мтДНК в байтах, что делает КДП запись предпочтительным видом хранения информации о мтДНК.
- 4) КДП характеристики мтДНК для живых существ, входящих в одну группу имеют характерные признаки, по которым можно однозначно определить принадлежность мтДНК к этой группе живых существ. То есть, каждое мтДНК хранит в себе КДП характеристики, однозначно определяющие его принадлежность к одной определённой группе живых существ.
- 5) Для однозначно определения принадлежности мтДНК к одной из групп живых существ, или просто идентификации мтДНК, необходимо последовательно применять КДП формулы (которые в статье

называются фильтры), достигая в каждом их применении выявления одной из КДП характеристик мтДНК последовательности.

Благодарности

Автор статьи искренне благодарит профессора МГУ, доктора биологических наук Троицкого Алексея Викторовича за оказанные им консультации по работе с интернет базой данных ДНК общедоступного пользования [7].

Библиографический список

1. Филатов О. В., Филатов И.О., статья **«О закономерностях структуры бинарной последовательности»**, «Журнал научных публикаций аспирантов и докторантов», 2014, №5 (95), с. 226 – 233.
2. Филатов О. В., статья **«Доказательство теоремы: «Формула для цуг из составных событий, образующих случайную бинарную последовательность»**», «Проблемы современной науки и образования», № 20 (102), 2017 г. с. 6-12.
3. Филатов О. В., статья **«Описание структур любых последовательностей образованных равновероятными случайными событиями»**, «Проблемы современной науки и образования», № 5 (138), 2019 г., с.9-15.
4. Филатов О. В., статья **«Описание распределения составных событий и их мизесовских частот через число возможных исходов. Механизм сжатия некоторых «не сжимаемых на один» последовательностей»**, «Проблемы современной науки и образования», № 9 (39), 2015 г., с. 27-36.
5. Филатов О. В., Филатов И.О., Макеева Л.Л. и др. **«Потоковая теория: из сайта в книгу»**. Москва, «Век информации», 2014, с.200.
6. Филатов О. В., Филатов И.О. **«Закономерность в выпадении монет – закон потоковой последовательности»**. Германия, Издательский Дом: LAPLAMBERT Academic Publishing, 2015, с. 268.
7. АДРЕСС БД ДНК: <https://www.ncbi.nlm.nih.gov/genome/browse#!/organelles>
8. Интернет ник автора: олегвладфилат.